THÈSE

présentée en vue de l'obtention du

DOCTORAT

DE

L'UNIVERSITÉ TOULOUSE I

Discipline: Mathématiques

Spécialité: Statistique

par

Sébastien Markley

Intégration de données spatiales dans la modélisation des choix discrets: applications aux modèles de comportements d'achats des ménages français

---

Soutenue le 28 Mars, 2008 devant le jury composé de

Mesdames, Messieurs les Professeurs:

Gérard Cliquet	Université de Rennes I	Rapporteur
Dominique Haughton	Bentley College (USA)	Rapporteur
Benoit Mulkay	Université de Montpellier I	Rapporteur
Jean-Philippe Lesne	BVA	Examinateur
Anne Ruiz-Gazen	Université de Toulouse I	Directrice de thèse
Michel Simioni	INRA	Directeur de thèse
Christine Thomas-Agnan	Université de Toulouse I	Examinatrice

Groupe de Recherche en Economie Mathématiques et Quantitative
UMR CNRS 5604, Université de Toulouse I, 21, allée de Brienne, 31000 Toulouse
Tel: +33(0)5 61 12 85 54

l'Institut National de la Recherche Agronomique
147 rue de l'université, 75338 Paris Cedex 07
Tel: +33(0)1 42 75 90 00

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Discipline: Mathematics

Specialty: Statistics

by

Sébastien Markley

Integration of spatial data into discrete choice models: Application to the modelling of French shopping behaviour

---

Defended on March 28, 2008 before the jury comprising

Gérard Cliquet	Université de Rennes I	Reviewer
Dominique Haughton	Bentley College (USA)	Reviewer
Benoit Mulkay	Université de Montpellier I	Reviewer
Jean-Philippe Lesne	BVA	Examiner
Anne Ruiz-Gazen	Université de Toulouse I	Thesis supervisor
Michel Simioni	INRA	Thesis supervisor
Christine Thomas-Agnan	Université de Toulouse I	Examiner

Groupe de Recherche en Economie Mathématiques et Quantitative
UMR CNRS 5604, Université de Toulouse I, 21, allée de Brienne, 31000 Toulouse
Tel: +33(0)5 61 12 85 54

l'Institut National de la Recherche Agronomique
147 rue de l'université, 75338 Paris Cedex 07
Tel: +33(0)1 42 75 90 00

L'université n'entend ni approuver ni désapprouver les opinions particulières du candidat

I would like warmly to thank my thesis supervisors, Anne Ruiz-Gazen and Michel Simioni for their guidance, their support, their suggestions, their corrections, and not least, for their infinite patience.

I would like to thank Jean-Philippe Lesne and Hervé Tranger at BVA for having proposed this project of study, for having supplied the data, for having provided my office and resources, and for the great help they have given me in learning and understanding the data and the models that we have developed here.

Due to the difficulties in bringing a private company and a public research institution into a three-way agreement with the French government agency ANRT as required by the CIFRE program, I am very grateful to my thesis supervisors, and my supervisors at BVA for their having been able to make my research project fruitful and interesting, satisfying on both a professional and intellectual level, and to have been unfailingly able to address my needs.

I would like to thank all those at BVA with whom I have worked with and enjoyed the company of, but in particular, Dina Likht with whom I have shared an office for three years now, and fellow CIFRE student Valérie Cerri. They have helped me on many occasions with the tedious work of proofreading and correcting my writing (especially what was written in French) checking the layouts of my presentations, and for that I am very grateful. I would also like to thank in particular two interns, Yoan Bolher and Sonia Barrieu, who both were assigned to a six-month work project with me, and provided me with both assistance and comradrie.

Finally, I would like to express my gratitude to Christophe Bontemps, Valérie Orozco, Céline Bonnet, and Ruxanda Berlinschi for having helped me prepare my thesis defense.

Contents

List of Figures

List of Tables

Introduction

A Frenchman does not need to be an expert on grocery distribution to have noticed that grocery retail has been changing in recent years. This is largely due to the aggressive introduction of more and more "hard-discount" stores in France whose distinctive and easily-recognised format consisting of no-frills store layout and service, extremely low prices, and restricted product selection has enjoyed a spectacular success. In just five years, from 2000 until 2005, the market share of hard discount stores in France has gone from 9 to 13.3 percent. At the same time, the percent of French households visiting hard discount scores rose to 66.8 from 55.3 percent between 2000 and 2004 (Leboucher, 2006 [23]). This has been a phenomenon dramatic enough to affect the daily lives of French people noticably and to attract media attention, and it serves to illustrate how the retail industry in France is fundamentally dynamic and competitive. This reality, when combined with the economic importance of food retail, generates a strong market for econometric analyses that will allow a greater understanding of the evolution and complexity of the agroalimentary sector, and the creation of the means to forecast future changes.

This is why the development of models of grocery consumption has attracted the interest of the private survey instute BVA in Toulouse[1]. This company has invested in

[1]The BVA Institute is a French survey institute founded in 1970 that in 2006 boasted 245 permanent employees and a gross revenue of 41 million euros. Before its recent restructuration, the company organized its expertise around four poles: surveys of popular opinion, measures and forecasts, marketing

the CIFRE Program[2] in order to create a set of econometric models that can be used as a tool to predict the number of clients who will visit any large-surface grocery store in France, real or hypothetical, and what products they will buy in these stores.

This project is anteceded by BVA's development of a model of spending by French households on various household products. Since the variable of interest (the amount of money spent by a given household on a given product) is always positive and is zero for a percent of the population (those not conducting any purchases), BVA uses a Tobit model to predict its value. The Tobit model is a censored regression model (described in, among other sources, Thomas, 2000 [45], pp. 121-155), that assumes that the amount of spending for each household is equal to a latent variable following a normal distribution when the value of this latent variable is positive, and is equal to zero otherwise. The estimation technique used by BVA is developed by Heckman (1976 [18]), and consists of two steps. In the first step, a binary Logit model of the probability of the household conducting a purchase (and spending being positive) is calculated. In the second step, an ordinary linear regression is calculated on only those individuals with nonzero responses. The bias of the estimate calculated over this subsample resulting from censoring is corrected by entering an adjustment term in the regression called the Mills Ratio which is based on the probability of obtaining a positive response calculated in the first step of the estimation. This allows BVA to produce unbiased predictions not only of the total amount of spending done by a population but also to forecast the

research, and geomarketing. Our research fell under the measures and forecasts pole. Headed by Jean-Philippe Lesne, Measures and Forecasts did studies of transportation networks, tourism, commerce and urbanity, environment and agriculture, and animal health.

[2]The CIFRE Program ("Convention Industrielle de Formation par la Recherche") is program administered by the government of France that finances public-private partnerships in research projects that yield doctoral theses. There are three partners involved in a project receiving CIFRE financing: a private company, a public research institution, and a qualifying student. The private company finances the student by hiring him/her for a three-year work contract and proposes the course of study. The public research institution works closely with the student on the development of the project and assures that the research results in a doctoral thesis.

number of individuals who make purchases, important information in its own right.

This model is used in the development of a geostatistical package provided over the internet that permits one to display at the click of a mouse interactive maps that show for each geographic subdivision of France the predicted probabilities of households purchasing each type of product and the predicted spending on the product. BVA is hoping to improve the information provided by this service by predicting not only how much money households from any given geographic sector of France will spend on each product on average, but how the money spent on each type of product is distributed over shopping locations.

For my doctoral thesis, BVA proposed that I pursue two related projects. In the first, I would develop a model of large-surface store choice from a data set generated in a survey of consumer behaviour undertaken in the Centre Region of France. This model would be used to predict the probabilities of households everywhere in France selecting each possible store choice. These probabilities could be summed by geographic zones to obtain predicted numbers of clients visiting each store coming from each zone. In the second, I would develop a model of the choice of type of product conditional on the model of the choice of large-surface store. In both cases, the model would have a predictive, rather than an explicative function. The ability of this model to forecast accurately the store and product choices of households in regions of France other than the Centre Region would be more important than the actual fit of our model to the data set we use.

In the long run, BVA has hoped to be able to create predictions of expected revenues of any store selling any product at any geographic point, and use this either to find the optimal store emplacements in France, or to evaluate store performance by comparing observed with expected revenues at each retail point. The final product presented to clients would not necessarily be the end result of our work. Rather, our project will

end with a collection of statistical procedures that managers at BVA can use over time, constantly integrating new data, and when necessary, modifying model structures so that the product remains up-to-date.

Little (1970 [24]) has developed the idea of a "Decision Support System" to describe management models that are used in an interactive way by managers as tools to aid in their decision-making. His idea is that for a management model to be useful, it is not only important for it to have good fit, or produce accurate predictions, but it needs to allow sufficient control and understanding by a manager that it can allow him/her to find the information necessary for his/her reasoning process. In a theory that has become very influential in management circles, Little theorizes that a successful decision support system ought to conform to six criteria: that it be simple, robust, easy to control, adaptive, complete on important issues, and easy to use for communication. Little goes on to say that a decision support system can be either a fully automated model, or a hands-on model. This conforms to two aspects of our research aims. We are in effect hoping to create a decision support system for investors in the large food industry, which Little would class as a "fully automated model". However, our project is also generating a type of decision-support system for the statisticians at BVA who will work on providing the final product to customers. The statistical procedures created here will be passed to them in a form that is most convenient for manipulation and understanding. In this case, the decision-support system is also a "hands-on" system that will allow managers at BVA to apply all their field and theoretical expertise to optimize their product.

The greatest focus of our project is on large-surface store choice, rather than product choice. There has been a lot of research done on the evolution of this phenomenon in France in recent decades, reviewed in Cliquet (2002 [9]), that can offer us insights. We know, for example, that once we have taken into account the very important dis-

tinctions between store types, the choice of grocery store essentially depends on travel cost. Economists divide goods into three types : commodity goods (goods that evoke no brand-loyalty, i.e. basic food products), shopping goods (goods whose purchase requires research and comparison of product characteristics, (i.e. cars) and specialty goods (goods whose attraction is strongly associated with brand, i.e. clothing). Households are more willing to tolerate travel costs in their research of different products when they select shopping goods, but this is not the case for the selection of food products. There is little variation in the inherent attractiveness of large-surface food stores apart from the admittedly very important division of stores into a few general formats. Where there is major variation in a household's consideration of different options, it is in the time and cost of accessing stores. Therefore, any model of store choice will have to take into account the ways in which stores are distributed in space, and the displacements habits of French households. Unfortunately, if models of shopping behaviour must reflect household displacements, then these models must adapt to the ever increasing complexity of French movements. Dion and Cliquet (2002, [10]) describe how French travel has changed over the last generation, explaining why simple measures of distance and travel time on their own (but not travel behaviour in general) are "disappointing" predictors of French store choice.

French households are travelling further on their displacements than before ([9]). Despite the fact that they did about as many trips in 1999 as they did in the mid-1970's and spent about as long on each of these trips (a total of about 55 minutes per day on average), greater transport efficiency has resulted in the average number of kilometers travelled per person per weekday rising by about a third to 23 km. Secondly, French travel is being done more often by car. In the last 15 years, the number of car uses in France increased by a third, while the number of households possessing two or more

cars has also increased. These factors cause household displacements to become more complex, as they occur over a larger area in which geographic disturbances are more likely.

A further complexifying factor is the way in which households are rationalizing their displacements ([9]), doing less one-stop shopping trips following straight there-back trajectories, and doing more multi-purpose shopping, following looped paths that hit several destinations on the same trip. This means that a French household will select stores more often out of opportunity because they happen to pass nearby it than simply because it is near their home or workplace.

French travel patterns also depend upon the purpose of the trip ([9]). It has been found, for example, that grocery shopping is done more often by car in France, while public transportation is more often used for social trips, and specialty item purchases, and more often used by the young and the elderly. The increased use of cars and the increased distances of trips has also been accompanied by a more chaotic organization of French displacements. These have become less predictable, as, probably due to a combination of traffic congestion and a more supple life schedule, households do not travel as often at set times, but choose travel times that are most convenient for them, avoiding rush hours, for example.

Two conclusions can be made from these developments in French behaviour. First of all, if our model of store choice is based on French shopping patterns, the validity of our model may be affected by the evolution of this phenomenon. It may be that proper use of this type of model will have to depend upon constant revision and reestimation of model parameters with fresh data. To address this concern, the model will need to be tested with data from different time periods. Since in this thesis, we study information collected only at a single point in time, the spring and summer of 2004, we are not

concerned with this. A second conclusion is that since French displacement patterns are complex, we need to use many variables besides distance in order to account for other factors that affect households' travel itineraries. We shall discuss how we have been able to make use of geographic variables detailing the population densities and accessibilities of the communes in which each large-surface store is located. For this reason our model of store choice will have a very important spatial component.

Contrary to what may be expected, sociodemographic variables do not play a determinate role in a household's decision. Due to the different shopping dynamics associated with each large-surface store model, we had initially believed that the choice of a type of store would be a matter of personal taste, influenced by social and psychological factors. This would mean that we could use sociodemographic variables to predict the choice of a store type (supermarket, hypermarket, etc.), and then use a second model to predict the choice of store based on the prediction of the store type and the spatial distribution of stores of this type. Unfortunately, we found that the sociodemographic variables were very poor predictors of choice of store type. Dion and Cliquet (2002 [10]) have confirmed that such variables do not suffice to predict decisions made by households since they do not account for the full variation in personal tastes. In this thesis, therefore, we present a model of store choice that we developed that does not depend on the characteristics of individual households, instead relying entirely on spatial and geographic effects.

Market share models based on explicitly spatial characteristics have a long history in econometrics. In 1929, Hotelling [20] developed a theory of the effect of travel cost on market share. A more explicit function of travel cost was introduced in Huff's market share model (Huff, 1963 [21]). According to this model, all else being equal, a store's market share will be proportional to its size and a negative exponential of its distance from the given population. The Multiplicative Interaction Model is an extention of Huff's

model that includes other variables beside size and distance whose parameter estimation through linear regressions on logs of observed market shares was developed by Nakanishi and Cooper (1974 [37]). This model continues to be popular in econometrics, and has been used recently by Cliquet (1995, [8]) to model the market shares of furniture stores.

Where data on store choices are available at the individual level, as in our case, market shares can be estimated by assigning probabilies of selecting each store to each individual, and then finding population totals for these probabilities. Doing this allows one to include explanatory variables in the choice probability calculations that are specific to individuals. It is a technique that requires the application of discrete choice models, the most well-known of which is the Conditional Logit Model developed by McFadden (1974a [32]). This model was first applied to travel modes on public transit (1974b [33]) and was applied in a marketing context in 1978 in a model of college choice by students (Punj, Staelin [38]). In 1981, it was used to model choices of large-surface stores in (Arnold, Roth, and Tigert [3]).

The continued use of discrete choice models has led to an exploration not only of the many variations of the known models such as Logit and Probit, but reflections upon how these models can be assumed to be accurate representations of behaviour. For consumers in particular, there has been considerable reflection upon how customers assimilate information in their shopping decisions. In 1975, Monroe and Guiltinan ([35]) developed a complex model that interrelated nine aspects of store choice that included

- Household/buyer characteristics
- Retailer strategies
- General opinions and attitudes concerning shopping
- Importance of store attributes,
- Perceptions of store attributes

- Attitudes toward stores
- Strategies for planning and budgeting
- Product and brand purchases
- In-store information processing

The relationships between these aspects were analyzed by applying this model to a local area whose supermarket retail was dominated by two outlets and observing the disturbances caused by the introduction of a third supermarket. More recently, Baker *et al.* (2002 [4]), modelled the environmental factors going into assessments of retail outlets with respect to patronage intentions. Their model included store choice criteria (interpersonal service quality perceptions, merchandise quality perceptions, monetary price perceptions, time/effort cost perceptions, psychic cost perceptions, and merchandise value perceptions) and a set of environmental factors (store employee perceptions, store design perceptions and store music perceptions). They posited a set of structural paths linking the different environmental factors with the household's assessment of the store according to the store choice criteria and the household's probability of making purchases in the store. By using path-analytical techniques to judge the fit of this model on a set of data, the authors were able to test long-standing assumptions about shopping behaviour.

Such reflections have been taken into account in discrete choice models of store selections, often in order to integrate spatial effects in decision hierarchies. In these models, the structure of decision-making translates into correlations between the likelihoods of different alternatives. If one uses Logit, these correlations can be incorporated through a generalization of the basic model structure. Fotheringham (1988 [14]) reasons that individuals are incapable of simultaneously comparing large numbers of stores and so their choices are done in a hierarchical manner according to the spatial configuration of the alternatives. Individuals will compare the combined utilities of groups that are clustered together geographically, and then compare the utilities of the stores within each group.

He suggests the use of the Nested Logit model when the geographic group, or nest, to which any store belongs is always known, and developes what he calls the "Competing Destinations Model" when it is not. More recently Guo ([16]) in her doctoral dissertation considers applying spatial hierarchies to models of residential location choice, but recognizes a difficulty in identifying geographic subgroups to which assumptions of proportional substitution can be applied. In these papers, there is a recognition both of decision-making being multi-staged and hierarchical, and an acknowledgement that the decision structures are not directly observable and must be accounted for by the introduction of unknown model parameters that are possibly costly in terms of computation.

The choice hierarchies described above do not factor in the effect of the increasingly multi-purpose nature of shopping trajectoires. In a multi-purpose shopping trip, these choice structures are less likely to be applicable to the household's choice of store, since the choice of store is motivated by the purposes of the trip overall, be it a home-work trajectory, for example, or a social trip, which may be independent of traditional criteria of large-surface store choices. Some reflection will be required on the implications of this on the choice behaviours that we attribute to households. Since the overall purposes of such displacements that include the large-surface shopping trips are unknown to us, we will make use of geographic variables that can act as proxies to indicate relatively likely household destinations.

Due to an increasing availability of spatially-oriented data, data sets containing precise geographic information at the level of the individual is more and more available and can be used to enter spatial effects as an explanatory variable in a discrete choice model, rather than as an effect that is approximated through the use of sophisticated model structures. This can be done in store choice models through the introduction of a distance variable representing the distances between individuals and large-surfaces stores

that can act as a proxy for travel cost. Smith (2004 [41]) uses a distance variable in his model of discrete choice of large-surface stores in order to study the effect of retail firm mergers on pricing. He begins by developing a discrete choice model of store choice that uses the spatial distribution of retail outlets and in which firm pricing strategies serve as explanatory variables. The price can then be estimated by assuming a Nash equilibrium and calculating the best fit of the prices to the known profits of the major grocery store firms. Doing this, Smith can model the effect of the merger of different store firms on prices. More recently, Turolla (2007 [48]) has also been working on the modelling of French grocery store choice. He has developed a combined model of multiple and ordered grocery store and food product choice in which he assumes that each household's choice of store for each product would be independent of the household's choice of store for other products. One fortunate quality of our research project is that we have access to quality geographic data that we can integrate into the explanatory variables used in our model. Unfortunately, integrating the effect of multi-purpose shopping on household behaviour increases the demands on our data, as records of home-work trajectories undertaken by households will likely be necessary to improve our model.

The decision process that we attribute to our households is not only a reflection of current thinking about behaviour patterns, but must also reflect the data that we use. The choices of stores made by the households in our survey are recorded using questions that follow a set pattern. Store choices are defined as being stores habitually visited by households, rather than a selection of stores made on a single purchase event. Purchases in large-surface stores are separated from purchases in traditional commerce. Store choices are ranked by households in terms of frequency of store visits, rather than amount spent on purchases in each store. Households are not asked to distinguish between store types (which would be difficult since most French people are not familiar with the

store types defined by retail specialists.) The model that we apply to this data will only indirectly refer to household behaviour since our data contains information about household behaviour as respondents fit it to the structure of the survey questionnaire. Our understanding of the logic of household store choices will have to reflect this fact. We have thus studied our model of store choice in a way reflecting the questions posed to each household. Since households were asked first which large-surface stores they visited, and then within each store, which very general categories of products they selected, we will create a model of household responses to the first question, and then a second model referring to the products chosen in each store that depends on the first. It may be surprising that household spending is not included in our model of store choice, nor is price included. This information was not included in our survey questionnaire, and is pricing strategies are not easily obtained from grocery retailers, due to the competitive nature of their business.

A model of store choice must not only take use spatial effects, but must also include differences between fundamental categories of large-surface stores. Large-surface food stores in France fall into three categories: supermarkets, hypermarkets, and hard discount stores. Supermarkets are defined as being large-surface grocery stores that have between 300 and 2500 square meters of retail space. They are smaller than hypermarkets, but are also far more numerous. These types of stores intend to attract local, regular shoppers, who would tend to make shorter, but more frequent shopping trips, buying fewer, and often more perishable products. Hypermarkets are defined as grocery stores having over 2500 square meters of retail space. These are far less numerous than supermarkets, but draw larger numbers of customers from a much larger area. Customers tend to do fewer shopping trips to hypermarkets, but buy more products. In our research, we split hypermarkets into two different categories: large hypermarkets, having over 8000 square

meters of retail space, and small hypermarkets. We believe that the largest large-surface stores have a different effect on customers, justifying a different treatment, since they are large enough to have transportation networks arranged around them and have the resources to maintain advertising campaigns that pull customers in from a great distance. As we noted in the first paragraph, Hard Discount stores are a large-surface store format that is still relatively new in France, but is rapidly expanding its market share. They are distinct from supermarkets and hypermarkets in that they provide much lower product variety, but undercut their competition with their pricing. They tend to be small in size, but very numerous, so as to be located as near as possible to their customers' homes, therefore minimizing their clients' travel burden. Hard Discount stores are identified as belonging to a brand that follows the hard discount business model. We will show in this thesis how important these distinctions between store types affect shopping behaviour.

Another aspect of grocery retail to be mentioned is the effect of branding, something looked at by Gonzales-Benito (1994 [15]) who studies the interaction between brand image and market segmentation. Unfortunately, our models cannot take into account the effects of branding on shopping behaviour, as other papers do, since we see pricing, marketing strategies, and store brands as specific to different regions, and therefore any model that is based on these variables is not useful for making forecasts in all of France. This information is also often proprietary, and we wish to be able to make forecasts based only on data that is publically available. Store choice hierarchies that include distinctions between store characteristics are sometimes used in store choice models. In his model of grocery store choice, Smith [42] supposes that households select a large store (one of the "big four" supermarket chains in England) and then selects a small supermarket as a function of the choice of large store for its food purchases. Smith therefore creates pairs of supermarkets consisting of on "big four" supermarket and one small supermarket

to which the Nested Logit model applies, the nests consisting of pairs of supermarkets sharing the same large supermarket.

Since our data is the starting point of our modeling, Chapter 1 begins with a detailed description of our data set, the sources of our data, and the variables we use. We use this context to develop more fully our reflections about shopping behaviour as we think it is reflected in the characteristics of our data collection. We also look at some exploratory statistics that give us a general idea about the behaviour that we observe. In Section 1.2, we introduce the modeling techniques that we use, beginning with a survey of the techniques described in the literature on studies of consumer behaviour. We describe the Conditional Logit model, and how it relates to Multiplicative Competitive Interation (MCI) models and gravitational models of store choice. We overview the technique, and its limitations with respect to our purposes, especially with regards to the computational resources required to run it. In Chapter 2, we describe how the model we introduced in Chapter 1 is modified in order to apply it to our data, and in particular, how we can define the choice set of large-surface stores from which each household chooses when the consideration set of each household is unknown. This chapter also discusses what indices we use to evaluate our model. Evaluating and validating our model cannot be done through more classic validation techniques such as those based on log-likelihood tests, since these are measures of a model's fit to a training set, and not tests of the accuracy of the forecasts based on this model, and so we develop our own measures of a model's validity. Due to the limitations of conditional logit models, alternative techniques such as a simplified gravitational model are explored, and we look at ways in which our conditional logit model can be modified in order to reduce computational burden. These techniques involve using a mixture of conditional logit modeling and the much simpler, yet quicker to calculate gravitational modeling.

When we wish to justify the use of a model in order to make forecasts, we must take into account the difference between two approaches to creating models of human behaviour. The first is a "model first" or non-conditional approach, in which we develop a model based on assumptions of human behaviour, before we apply them to a particular set of data. An advantage of this approach is that the model will be universally applicable, as it is not premised on any individual case. The second approach is a data-driven approach. We investigate an actual set of observations, and then we build our model up from what we discover in our data. This is by definition a more empirical approach than the former and enables one to take into account the particularities of our case at hand. However, one disadvantage is that it is constrained by the data upon which it is based. The application of a nonconditional model to a particular case will only be valid insofar as the case conforms to the assumptions of our model. The application of a conditional model, on the other hand, will only be valid to the degree to which the data in our particular case resembles the data set on which the model is based. In our thesis, we follow both approaches to some extent. In the beginning, we develop a model of store choice through a priori assumptions of human behaviour. However, the structure of this model is also adapted to our data. The a priori and the a posteriori parts of our model development correspond roughly to Chapters 1 and 2 in our thesis.

In Chapter 3, we compare and evaluate the estimations of our model, using the evaluative criteria that we describe in the previous chapter. There already exists a previous work (Markley, 2006 [31]) in which the results of the estimations of this type of model are discussed and illustrated on a subset of our survey region. Chapter 4 of our thesis is consecrated to the attribution of product categories to the choices of large-surface stores by households in France. This is done without using a modelling strategy, but through a type of imputation. In this chapter, we discuss different imputation techniques

that we use, and the different ways to apply the data that we have at our disposal for these imputations.

Chapter 1

Data Description and Modelling Strategy

In this chapter, we introduce the data set provided by BVA that we use in order to create a model of store choice. This is based on a survey of household spending patterns in the Centre Region of France. In Section 1.1, we describe how we obtain the data that we use in our modelling procedures: the sampling strategy used in the survey, the behaviour recorded during the interviews, and the auxiliary information collected on the households. We also discuss the characteristics of our sample and the Centre Region in which it is taken from. Once we have introduced the data, in Section 1.2, we discuss Discrete Choice Theory and the theoretical background of the Conditional Logit model that we will use in Chapter 3 in order to predict the probabilities that households will select each large-surface store.

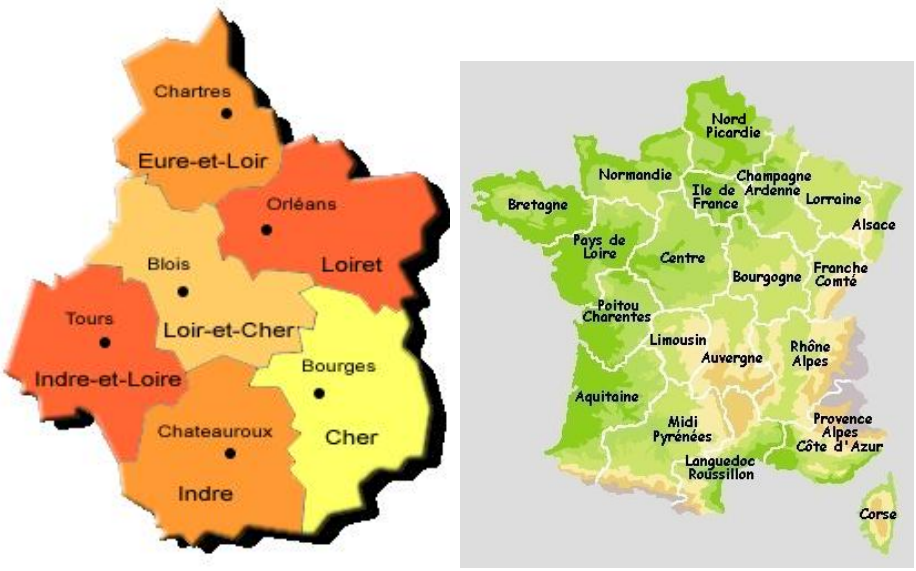

Figure 1.1: The Centre Region of France.

1.1 The BVA survey of household shopping flux

The data we use comes from a survey of households in the Centre Region of France undertaken by BVA in the spring of 2004. The purpose of this survey was to calculate shopping flux between customers of this region and destinations inside and outside the region. This data was used in order to create analytical reports containing, among other things, catchment areas of cities within the region. The study was commissioned by the chambres of commerce of the departments of the Centre Region, and it covered semi-durable as well as grocery commerce.

For the survey, interviews were sought with a member of each of the 14,217 respondent households selected for the survey sample through stratified quota sampling. During these interviews, a questionnaire was filled out in which the representative individual would provide information on the socio-demographic characteristics of his/her household and the geographic co-ordinates of the household's domicile as well as details on the household's shopping behaviour. Among the questions asked was which three large-

surface grocery stores the household visited most often for food purchases, in ranked order.

1.1.1 Sampling strategy

The Centre (or Val-de-Loire) Region of France is one of the 22 administrative regions of France. It is located just southwest of Paris, and is comprised of six deparments. We refer to the departments by the numbers assigned to them in French administrative records. These are: the Cher Department (18), the Eure-et-Loir Department (28), the Indre Department (36), the Indre-et-Loire Department (37)[1], the Loir-et-Cher Department (41), and the Loiret department (45). These are shown on Figure 1.1

Our survey region actually excludes the Loiret department, since it was studied in another survey. Each of these departments is in general centred on one large city that dominates the department. The 1999 census population of the Centre Region was 2,440,329 people, making up about 4 percent of the population of France. Table 1.2 shows the populations of all the departments in the region along with the population of the major urban agglomerations of the region.

The survey sample is selected from the households of the five departments included in our survey through quota sampling by strata. BVA defines the strata by dividing the survey region into geographic sectors, and segmenting the population by household type. The geographic sectors developed for the purposes of the survey are defined so as to be roughly homogeneous in terms of household characteristics, and have little variation in population. In the Centre Region, as seen in Figure 1.2, the resulting sectors range in size from 1080 to 10100 households, with three-quarters of sector populations containing between 2200 and 4100 households.

[1]Not a spelling error. The Loir, the Loire, and the Loiret are three different rivers

Department	Population	Urban Unit	Population
Eure-et-Loir	407,665	Chartres	88,318
		Dreux	44,663
Loir-et-Cher	314,968	Blois	65,989
Indre-et-Loire	554,003	Tours	297,631
Indre	231,139	Chateauroux	66,082
Cher	314,428	Bourges	91,434
		Vierzon	32,528
Loiret[2]	618,126	Orleans	263,292
		Montargis	53,590

Table 1.2: Populations of the departments of the Centre Region and their major cities (Urban Units) according to the 1999 Census of France

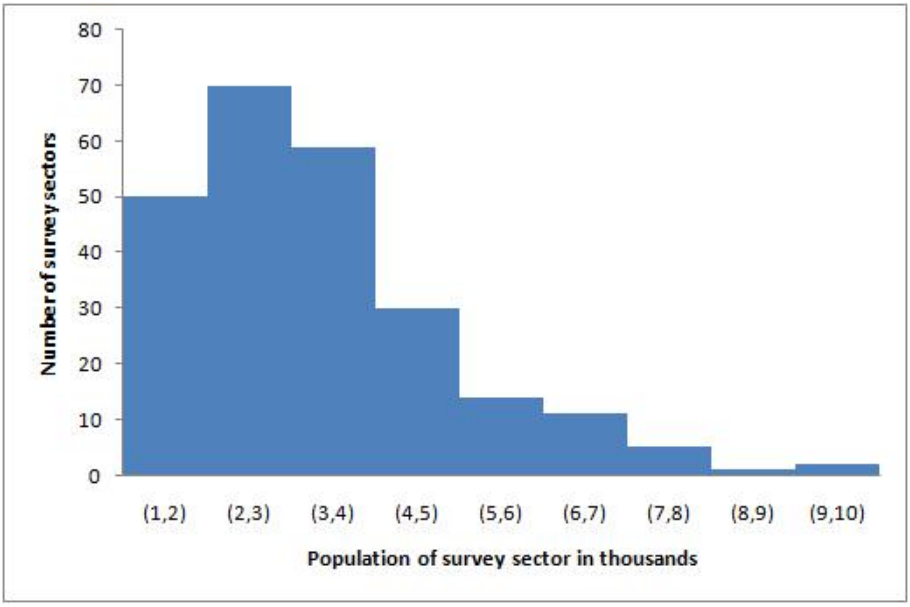

Figure 1.2: Survey sector populations for the BVA survey of shopping flux.

The survey sectors are defined as groupings of contiguous "IRIS"[3], which are geographic units defined by INSEE[4] for use in the national census it administers. These IRIS are partly based on communes, which are the smallest class of administrative district in France and are the geographic zones represented by municipal governments. Most communes are rural, with small populations, but some represent large cities, and thus must be divided into smaller geographic zones for the purposes of data collection. With every census, INSEE, in co-operation with commune administrators, divides the French territory into geographic zones called "Ilots" that are determined by the features of the land, and then for the census in question, aggregates these ilots in order to form continuous geographic zones termed IRIS. In the design of the survey, the BVA Institute groups IRIS in order to form survey sectors that are as homogeneous as possible, while representing a population of a reasonable size. In rural areas, sectors cover large areas, and group several communes, while communes representing cities are divided into several sectors. This means that although sector populations have reduced variation, they do vary a great deal in terms of surface area and population density.

In Figure 1.3 we show a map of the survey sectors defined by BVA over the survey region defined for its study of shopping flux. Red lines indicate department boundaries, and blue lines indicate sector boundaries. Within each of the large cities, survey sectors are numerous and very small, so they are not visible on this map. We recall that the Loiret department in the northeast is not included in our survey area, although BVA did divide this department into sectors for the purposes of another survey. Survey sectors are also defined for areas outside the Centre Region, although all the households in our datasets are within the survey region. The points represent large-surface stores, red being hypermarkets, either large or small, green representing hard discount stores and

[3] *Ilots regroupés pour l'information statistique*, or "aggregation of zones for statistical information".
[4] *Institut National de la Statistique et des Études Économiques*

purple representing supermarkets. In order to assure that all stores that are considered by consumers in the Centre Region are entered in our data, we include large-surface stores in all departments bordering our survey area.

Within each survey sector, a representative sample is selected using quotas defined for five different classes of households: households headed by a single employed person, households headed by a single unemployed person, households headed by a couple in which both individuals are unemployed, households headed by a couple in which only one individual is unemployed, and households in which neither individual is unemployed. We make great use of these survey sectors when we analyze our data since they are the smallest geographic units at which the sample is designed to be representative.

It is important to note that our sample is exogenous. That is, the sample comes from outside the model, and is not explained by the model. The selection of individuals in the sample depends upon their attributes, and not upon their observed responses. An example of a sampling strategy in discrete choice theory that is not exogenous would be a choice-based sample, where individuals are chosen on the basis of the choices they have made. The importance of this will come into play in Section 1.2.2 when we look at whether we should weight the observations in our data set in order to take into account the effect of sampling on the model estimation.

1.1.2 Determining store choice

An important feature of the survey is that the choices of large-surface grocery stores listed by each household are identified and geolocalized. The Chambres of Commerce of the departments of our study have provided a list of large-surface grocery stores contained within the survey region. The households' declared choice of store can therefore be matched to a store on this predefined list. Stores that are not on this list (mostly stores

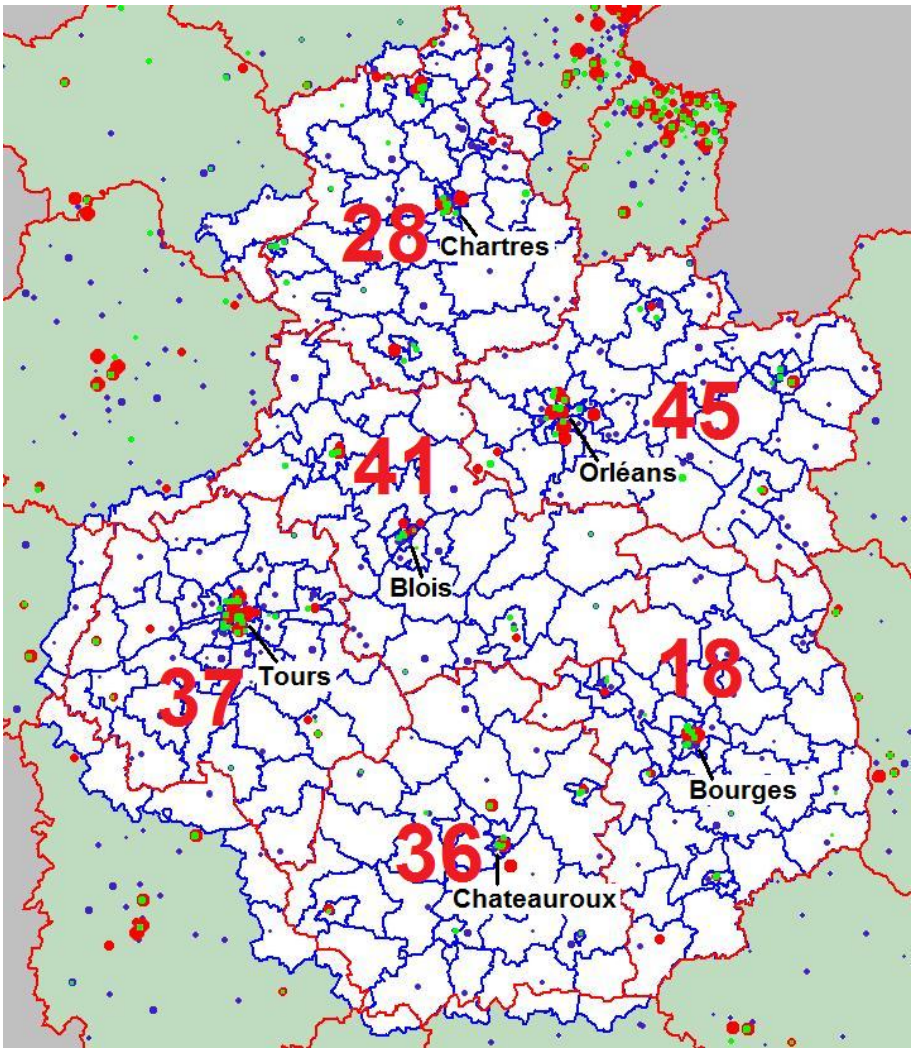

Figure 1.3: Map of survey sectors in the Centre Region of France along with large-surface stores. (Red dots represent hypermarkets large and small, green dots represent hard discount stores and purple dots represent supermarkets)

that are outside the region of study) are added as they appear in survey responses. Later, this list of stores is checked for inconsistencies. Erroneous entries for large-surface grocery stores are corrected or eliminated, and stores added to the list that are found to be duplicates of other stores already on the list are eliminated. For each store in the file, we record the location and geographic co-ordinates of the store, the retail space, and the store firm. Because we need to obtain a complete record of stores that are considered within the consumers' choice of store, and we know that households in the Centre Region frequently visit stores outside the region, we use a data set of large-surface stores in France purchased by BVA in order to add all stores contained within all departments bordering the Centre Region to our data.

The nature of the survey questions imposes a structure to household's stated shopping behaviour. Households are required to separate their grocery purchases in large-surface stores from their purchases in traditional commerce, they are required to order their choices of by frequencies of store visits, and they are required to separate their products purchased in one large-surface store from the others chosen. We have reflected on different aspects of a possible choice hierarchy, included:

- Number of stores chosen
- Distribution of visits among stores
- Store types (supermarket, hypermarket, etc.)

We wished to see whether perhaps households first selected a shopping pattern, perhaps in line with their socio-demographic characteristics, and then would select their three choices of large-surface stores based on this shopping pattern. We attempted dividing the population into groups according to their three choices of large-surface stores. For example, we created categories of households with respect to the percent of visits

for each household done in one large-surface store, by the types of stores visited by each household, by the number of stores the household visited (1,2, or 3). There was no evidence that households selected their shopping behaviour prior to selecting their store choice, since the groupings defined above were not determined by sociodemographic criteria. As well, the criteria of shopping behaviour were difficult to define since the order of store choices did not reflect the store's importance to the household, since the number of visits to any store did not at all reflect the amount of money spent at each store. The data seems to suggest that the number of stores a household visited, or the relative number of visits made to each store within those chosen by any given household is more a reflection of the spatial distribution of stores around the household than of the household's characteristics.

We have chosen a modelling strategy in which we have applied conditional logit models to each of the three choices of large-surface store for each household taken separately. This has the implicit assumption that they are done independently by each household, and that the qualities of all large-surface stores are simultaneously compared. An advantage of applying such a model is its simplicity and practicality due to low computation times and ease of use.

1.1.3 Exploratory statistics

In doing some exploratory statistics of our sample, we have investigated the primordial role that store format and store distance play in store choice. Any application of discrete choice modelling to large-surface store choice is necessarily influenced by store type. Indeed empirical evidence suggests that consumer choices of large-surface shopping stores depend greatly on whether they are supermarkets, hypermarkets, or hard discount stores. Supermarkets have between 299 and 2500 square meters of retail space, hypermarkets

have 2500 or more square meters of retail space, and hard discount stores are a particular category of supermarket, usually between 300 and 800 square meters, characterized by a smaller range of goods offered, but at a very low price. We include a histogram showing the distribution of these stores by their surface areas.

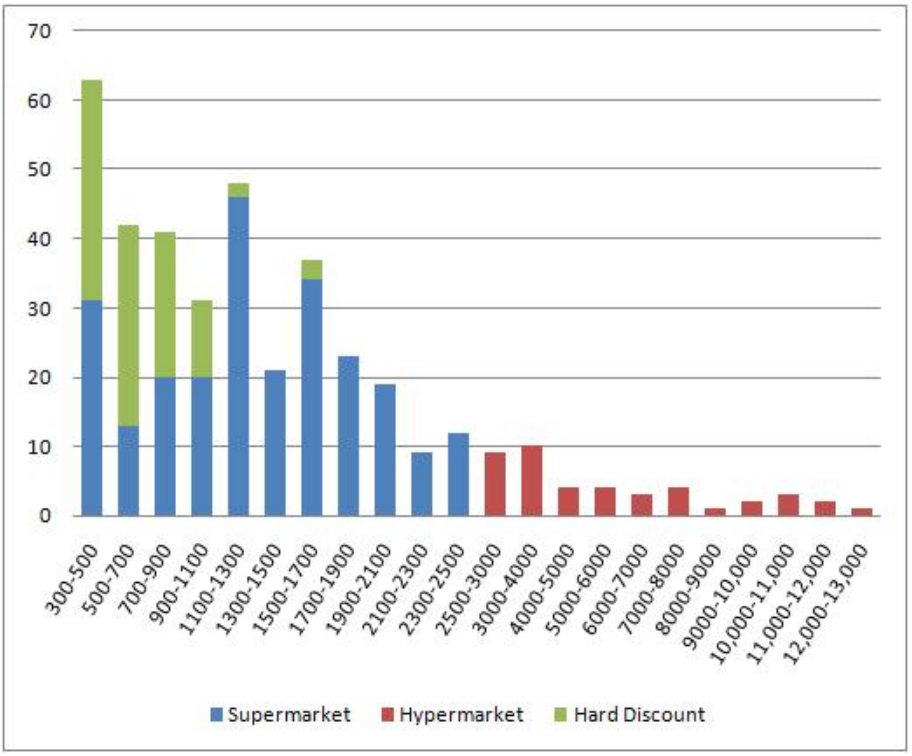

Figure 1.4: Histogram of stores in the Centre Region by type and retail space.

We believe that the hypermarket category contains too great a range of different store types, and that this class can be broken further down into large and small hypermarkets that would exercise categorically different effects on households. The experts at BVA indicated that they wished to identify the few very large hypermarkets that dominate the market in each department, that is, those hypermarkets that would be the largest hypermarkets that a household in its catchment area would consider.

In order to find where we can draw such a distinction, we divide hypermarkets into different categories, each containing retail spaces that differ by no more than 1000 square meters, and then looking at the average number of households in our sample visiting each store per square meter. This is shown in Figure 1.5. Smaller hypermarkets are small enough that they don't need to attract customers from a great distance in order to obtain a high client-to store surface ratio. We see that in our graph, as we move towards larger categories of stores, there is a dip in the bars at the 8000-9000 square meter category, where it seems increasing the size of the store does not increase the size of the store's catchment area enough to increase the number of clients at the same rate. However, as the store increases further in size, it seems to pass into a new category, since despite having a larger retail space, the number of clients with respect to its retail space is even greater. This seems to be a good sign of where a large hypermarket achieves market-dominating size. In all, there are 9 large hypermarkets in our survey region.

We therefore divide the category of hypermarkets into hypermarkets of retail space less than 8000 square meters, and retail space greater than 8000 square meters. In the rest of this document, we refer to supermarkets as "SM", small hypermarkets as "HM", hard discount stores as "HD", and large hypermarkets (over 8000 square meters) as "XM".

In Figure 1.6, we look at the proportion of households selecting each type of store. We note that all but 2 percent of households choose at least one large-surface store for its shopping needs, 75 percent of the population choose two or more and only 31 percent choose three stores. The charts show the percent of households choosing each type of large-surface store for the first, second, and third choice conditional on there being a store visit (nonchoices are not counted).

We see that households choose supermarkets much more often for their first choice of store than for their second and third. As we go from the first to third choice of large-

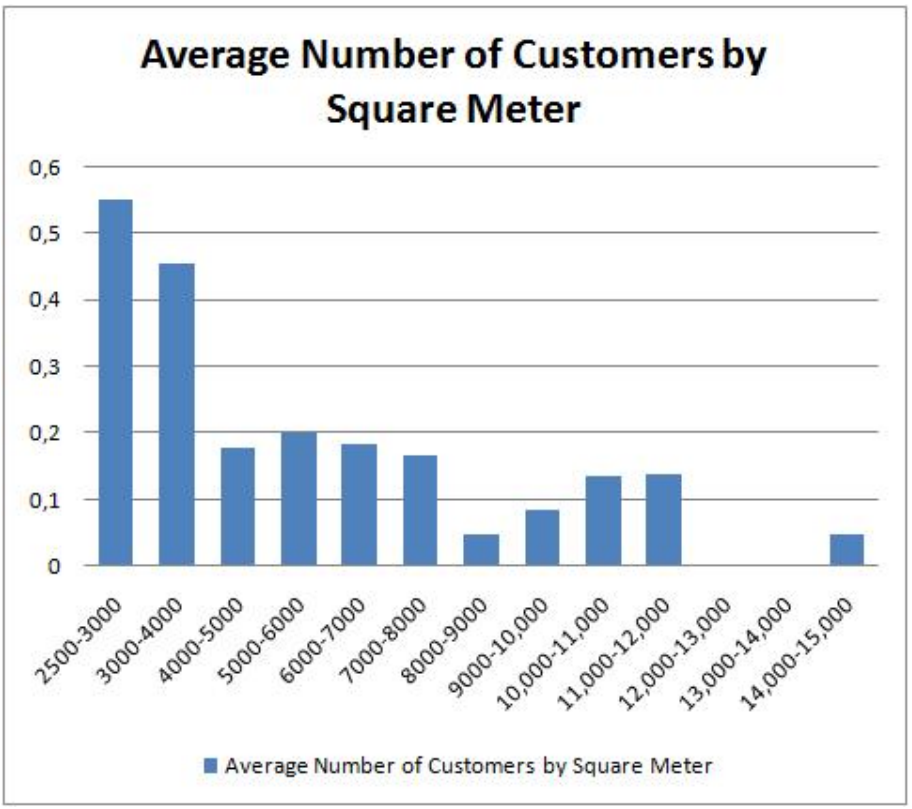

Figure 1.5: Average number of customers per square meter of retail space for hypermarkets by category of hypermarket size.

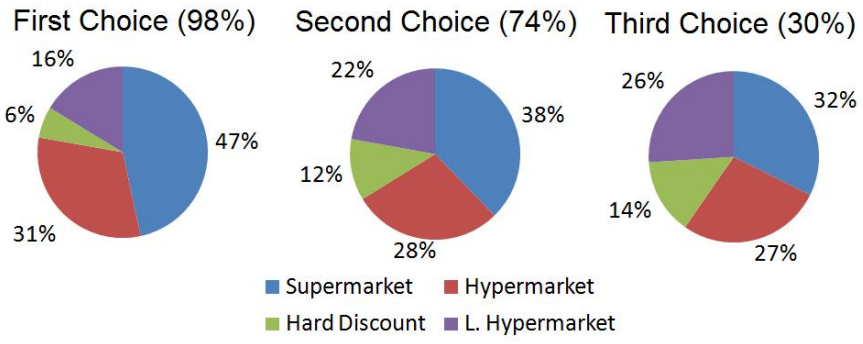

Figure 1.6: Proportion of households selecting each type of store by order of store choice.

surface store, households more often choose hard discount stores and large hypermarkets, and less often supermarkets and small hypermarkets. This corresponds to the tendency of many French households to visit supermarkets frequently in order to buy perishable goods, or to fulfill their immediate needs, and then to go to hypermarkets and hard discount stores once in a while to buy as many goods as they can in a single trip and at a lower price.

We would expect that households would tend to minimize the cost of a shopping trip, so we would expect them to choose the closest store to its residence, all other factors being equal. And indeed 22 percent of the time a store is cited as one of the three choices of large-surface stores for a household, it is the closest to the household's co-ordinates (the centroid of the household's IRIS of residence). We also see that a choice of supermarket ("SM") is far more likely to correspond to the closest large-surface store to the household, than a choice of hypermarket large or small ("HM" and "XM", respectively) or hard discount store ("HD"). Table 1.3, shows the percentage of each category corresponding to the closest (or tied for closest), second-closest, and third-closest store to the household's co-ordinates. For example, we see in the first line of the column under "SM" that 42 percent of the times a household chooses a supermarket for one of its three choices of stores, it is the closest large-surface store to the household's home. However, only five percent of choices of large hypermarkets correspond to the closest large-surface store to the household. This means that the effect of the rank of the distance on a household's choice depends greatly on the type of store the household considers.

Hypermarkets, especially large hypermarkets, are designed to draw households away from their homes, providing an appeal and a convenience that outweighs their distance. Thus, a close supermarket is not necessarily more attractive to a household than a far

	SM	HM	HD	XM	Total
Closest	42	9	15	5	22
Second Closest	16	9	11	5	11
Third Closest	10	7	9	3	8
Other Stores	31	75	65	87	59
Total	100	100	100	100	100
Pct of all choices	41	30	9	20	100

Table 1.3: Percent of store choices corresponding to closest store to domicile

hypermarket. Once we take into account the choice of store type made by the household, the effect of the rank of the distance of the store becomes far more clear. In Table 1.4, we look at the proportion of choices of each type of large-surface store that corresponds to the closest large-surface store within its category. This shows us that 50 percent of the time a household chooses a supermarket for one of its three choices, it is the closest supermarket to the household's home co-ordinates. We see now that over half of the time a household chooses a small hypermarket, it is the closest small hypermarket to its home, and over half the time a household chooses a large hypermarket, it is the closest large hypermarket to the household's home. This behaviour pattern seems to be less well-maintained for hard discount stores, however. This may indicate that the choice of hard discount stores is less sensitive to distance, and perhaps sensitive to other factors, such as brand recognition, or familiarity.

Besides showing us that people frequently shop in nearby stores, knowing that there are 1600 stores in each household's choice set, these tables show us that the probability of selection for most stores is extremely small and little will be gained by having a model that attempted to predict it accurately. As well, estimating a model can become more difficult if the introduction of the utilities of these very unlikely events will increase the uncertainty of our model estimation by introducing a very large number of degrees of

	SM	HM	HD	XM	Total
Closest	50	53	39	63	52
Second Closest	18	21	21	24	20
Third Closest	10	8	12	6	9
Other Stores	22	18	28	7	19
Total	100	100	100	100	100
Pct of all choices	41	30	9	20	100

Table 1.4: Percent of store choices corresponding to closest store to domicile conditional on type of store

freedom.

The most important data that we collect in our survey are the geographic co-ordinates of the household's home and the stores listed in our survey area, which enable us to create a data set that contains the Euclidean distance between each household in our sample, and each large-surface store in our file, and more importantly, determine which large-surface store is the closest to each household. This kind of information is in general very expensive, and is not often available to those studying shopping behaviour.

Unfortunately, although we record the addresses of the households interviewed in our survey, the cost of transforming addresses into exact geographic co-ordinates is far too expensive to be done. We therefore take as the co-ordinates of the household's home the centroid of the IRIS of residence, or the center of mass of the population of the IRIS in cases where this corresponds to a single commune. This obviously means that many households are assigned the exact same geographic co-ordinates. The imprecision of the co-ordinates of the households could be a source of model error, especially if it is great enough to cause us to be mistaken about the store that is closest to the household. However, we believe that although it must be admitted as a source of error, the IRIS are a very fine geographic definition. In urban areas, they represent a very small area, and in rural areas, store locations are more spread-out, making geographic precision less

necessary.

The geographic-co-ordinates of the stores in our survey, on the other hand, are more precise, corresponding to the centroid of a polygon drawn around the commercial zone in which the store is located. Despite this greater precision, many stores are assigned the same co-ordinates as their neighbours.

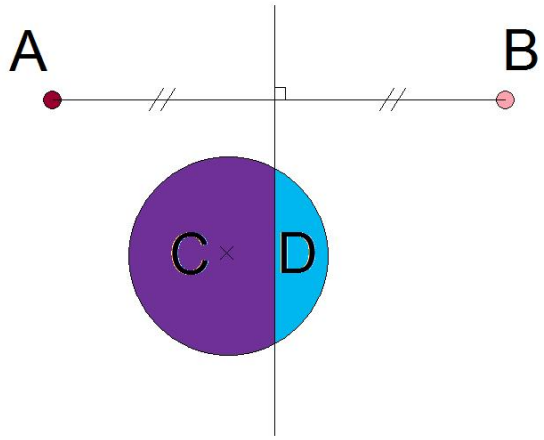

Figure 1.7: Estimating the effect of the imprecision of household co-ordinates

Due to the demonstrated importance of the effect of being the closest store to a given home, we need to see to what extent our imprecision leads us to be mistaken about what stores are nearer a household's home than others. In order to quantify this, we begin by assuming that all IRIS are exactly circular and their populations are spread uniformly across their surfaces. We then calculate the probability that each household, if it were assigned a geographic co-ordinate drawn randomly from within its IRIS, would be closer to the second-closest store of a given type to the attributed co-ordinates of the household than the closest store of the same type. In the cases where the two stores are in the same location, we assign a probability of 0.50. Taking the sum of these probabilities will give us a rough estimate of the expected number of households, if assigned the true geographic

co-ordinates of their homes that would have the closest and second-closest large-surface stores in a different order than with the current, less-accurate co-ordinates. In Figure 1.7, we illustrate how we calculate this. If the circle in the diagram represents the IRIS, "A" and "B" represent the two stores of a given store type closest to the co-ordinates of the IRIS, which is the centre of the circle. If any household is located within the area "C", then it will be closer to "A" than to "B", and the assigned rank of the distance of both stores would be correct. If a household were located within the area "D", then the household would be closer to "B" than to "A", even though in our data, store "A" would still be considered closer to the household. Thus, area "D" would be the zone of error. Assuming all households were equally distributed over the IRIS, the proportion of the circle covered by "C" would be equal to the probability that a household, drawn at random from the IRIS would be assigned the correct ranks of store distances. Calculating this probability for each IRIS, and then taking the sum weighted by the 1999 populations of each IRIS, we obtain the values presented in Table 1.5.

Store Type	% error
Supermarkets	%18.6
S. Hypermarkets	%6.2
Hard Discount	%12.5
L. Hypermarkets	%3.4

Table 1.5: Sensitivity of rank of store distance to accuracy of co-ordinates assigned to each household.

We believe these values to be somewhat pessimistic, for they ignore the effect of having populations concentrated in one part of the IRIS, as in the case of a village, contained within a rural IRIS, which would increase the probability that a randomly selected household's location would be closer to the geographic co-ordinates assigned to the household. However, replacing the complex polygons defining each IRIS with a circle of the same area will also reduce the probability of a false assignment of distance

ranks. From our results, we can assume that rendering our geographic co-ordinates more accurate would have very little effect on the correct determination of the closest supermarkets, and even less on the determination of the closest hypermarkets.

1.1.4 Sources of explanatory variables

So far, in our exploration of our data set, we have determined that households tend to choose the closest store within its category of large-surface store. We have at our disposal a great deal of information that we could use in order to predict the choice of large-surface store by individual. There are three categories of variables that we use to represent this information: variables referring to the socio-demographic characteristics of the households in question, variables referring to the characteristics of the large-surface store and its distance from the household, and finally, variables referring to the characteristics of French communes. In the first category, we have variables such as household size, income, and access to transportation in addition to the characteristics of individuals within the household, such as age, sex, and employment. In the second category, we have the distance of the large-surface store from the household, its category (supermarket, hypermarket, etc), its surface area, and its advertising logo. In the third category of variables, we have commune characteristics such as population, polarity, access to major highways, etc. A priori, we can say that a household will be more likely to visit a store not only near its home, but in an area household members tend to go for work, study, or leisure. We do not have direct information on where household members work, go to school, or pass their evenings, but our geographic variables can help identify the areas that are more likely to attract household members for these purposes.

Our first source of explanatory variables is our survey. Our respondents provide detailed information on the socio-demographic characteristics of the households. We

have the age, sex, socio-professional category, and primary means of transportation used in commutes for each individual in each household. We also have the total revenue of the household, and the type of residence in which it resides. We represent this information using the following variables:

- Quota category: Household headed by single employed person, single unemployed person, couple with both employed, couple with only one person employed, and couple with neither member employed
- Access to vehicle
- Presence of children in the household
- Number of people in the household: 1, 2, or more than 2
- Revenue category of household
- Residential category: Single residence, or multi-household lodging
- Household is proprietor or rents
- Age category
- Employment
- Socio-professional category: unemployed, profession requiring at least a university degree, and profession not requiring a university degree

Aside from the data provided directly by the survey respondents, there were also the variables that represent the qualities of each large-surface store with respect to the household in question. These include:

- Store Type: Supermarket, Small Hypermarket, Hard Discount, Large Hypermarket
- Retail space
- Distance of store from household domicile
- Store trademark

Aside from the information provided by survey respondents, we think it is important to incorporate information on the geographic characteristics of each commune in the survey region. These could be entered into our model twice, as qualities of a household (being the characteristics of the commune in which the household lives) and as qualitites of the large-surface store (being the characteristics of the commune in which the large-surface store is located). The data we used were from publically available sources compiled by INSEE, including:

- The census responses in an INSEE sample of 1/20 of the 1999 French census respondents
- The commutes done between every pair of communes in France by type of transportation, as recorded by the 1999 French census
- The 1998 Inventory of Communes in France

We use a sample of 1/20 of the population of France provided by INSEE that is representative of the population at the IRIS level. This sample contains all information entered by the respondents in the 1999 French census, except information that identifies them. What we find useful is that this data set also contains the INSEE classification of communes in terms of polarity or centrality. The inventory of French communes contains information on the types of infrastructures and services in each commune, as well as containing geographic information, such as the time it takes for inhabitants of the commune to access the nearest highway, or the nearest neighbouring town or city.

The characteristics of each commune that we use as explanatory variables in our models include the following:

- "Polarity" of commune
- Commune as city centre or as periphery
- Traffic between commune A to commune B by type of transportation

- Population of commune
- Density of population of commune
- Surface area of commune

The term “polarity” is a term developed by INSEE which requires a little additional explanation. Essentially, this is a classification of communes into four types in terms of their economic centrality:

- An urban pole is a set of contiguous communes in which at least 5000 jobs are located and that contains the workplaces of residents of surrounding communes.
- Monopolarized communes are not in urban poles but their residents tend to work in one urban pole
- Multipolarized communes are not in urban poles and not monopolarized but their residents tend to work in several urban poles
- Nonpolarized communes have residents who don’t tend to work in any urban poles.

This concept of economic dependence between communes is also an indicator of travel between communes. Since by definition, people in monopolarized communes travel more often to urban poles to work, we assume that they will logically be more likely to shop there as well. As we can see in Figure 1.8, the departments of the Centre region have "bull’s eye" formations in which communes go from being monopolarized to rural as they are further from the economic heart of the department. We can also see in the Northeast of the map where there is a mass of communes that are monopolarized since they are within the economic footprint of Paris. When two urban poles are near each other, there are grey zones where the economic pull of the two urban balance each other, and households are evenly divided by those who work in one urban pole, and those who work in another. None of the other urban units in the region have more than 11,000 inhabitants, therefore they are too small to have the requisite 5000 jobs in order to be considered an urban pole.

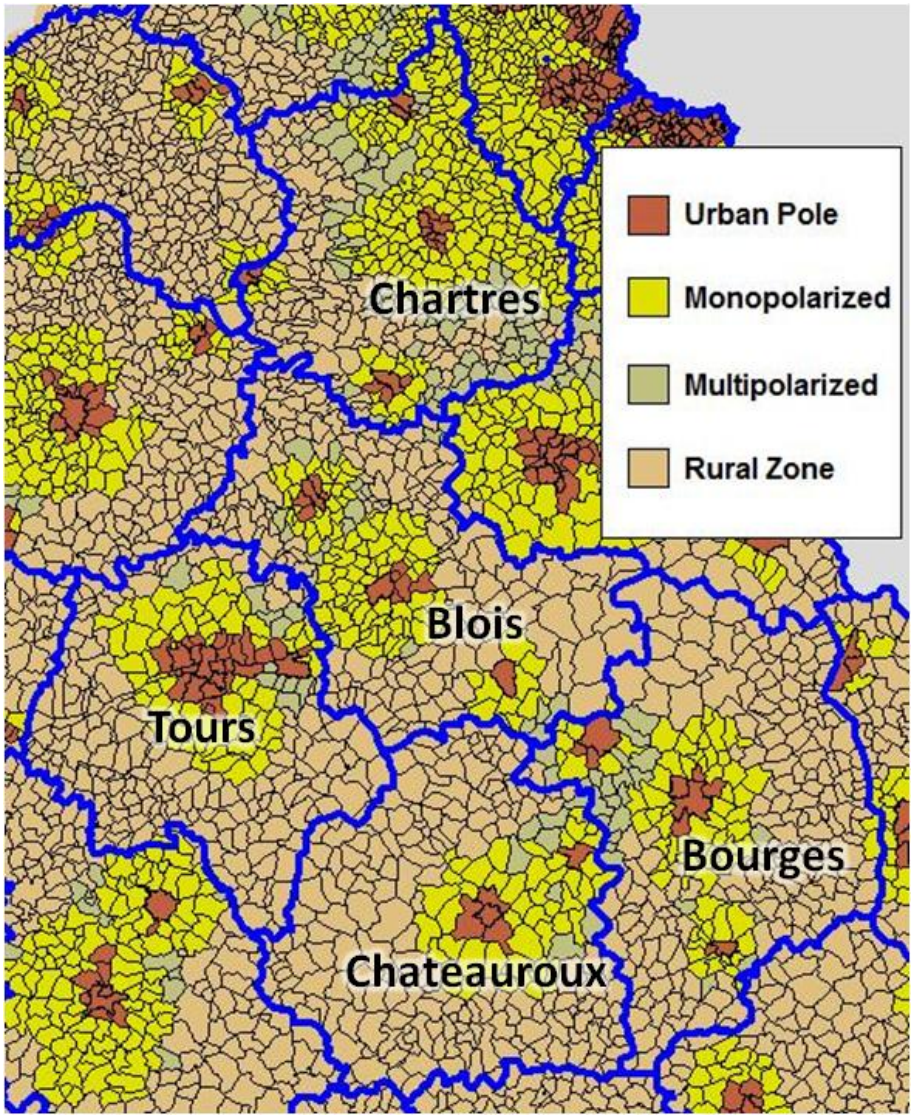

Figure 1.8: Polarities of communes in the Centre Region.

The polarity of communes shows the interactions between different settlements, but does not distinguish between the centrality of different communes within the same urban unit. This is why another INSEE variable also categorizes communes in France by their centrality. Inner-city communes are those communes in an urban unit containing at least 50 percent of the population of the urban unit, or having a population greater than 50 percent of the population of the most populous commune in the urban unit. Communes in an urban unit, but not inner-city communes, are considered suburban communes. The remaining communes have less than 2000 inhabitants, are not in urban units and are therefore classed as rural communes. Figure 1.9 shows where there are clusters of populations. We see some city centres that are on the edge of the major urban units, representing small towns that are near Tours, but not yet completely engulfed by the city's expansion, and therefore not considered as part of the Tours urban unit. Bléré, Amboise, and Chinon are all urban units comprising several population centres, and whose population is not concentrated in one of its constituent communes.

These census classes of communes will be taken into account in our model in order to rate the attractiveness of large-surface stores contained within them. We see by the preceding maps that we do not need to worry much about the effect the definition of the boundary of the survey area will have on our model. The area that is influenced by Tours appears to be completely contained within the Indre-et-Loire department, while at the same time, the communes depending on outside communes such as Blois, Saumur, and Châtellerault, are kept outside the survey area.

We can compare the distribution of the polarities and the centralities of the communes within the survey area with the observed road traffic in the same area shown in Figure 1.10. This map shows the average number of vehicles per day on the different parts of the road network connecting the major centres of the Centre Region in 2004, the year

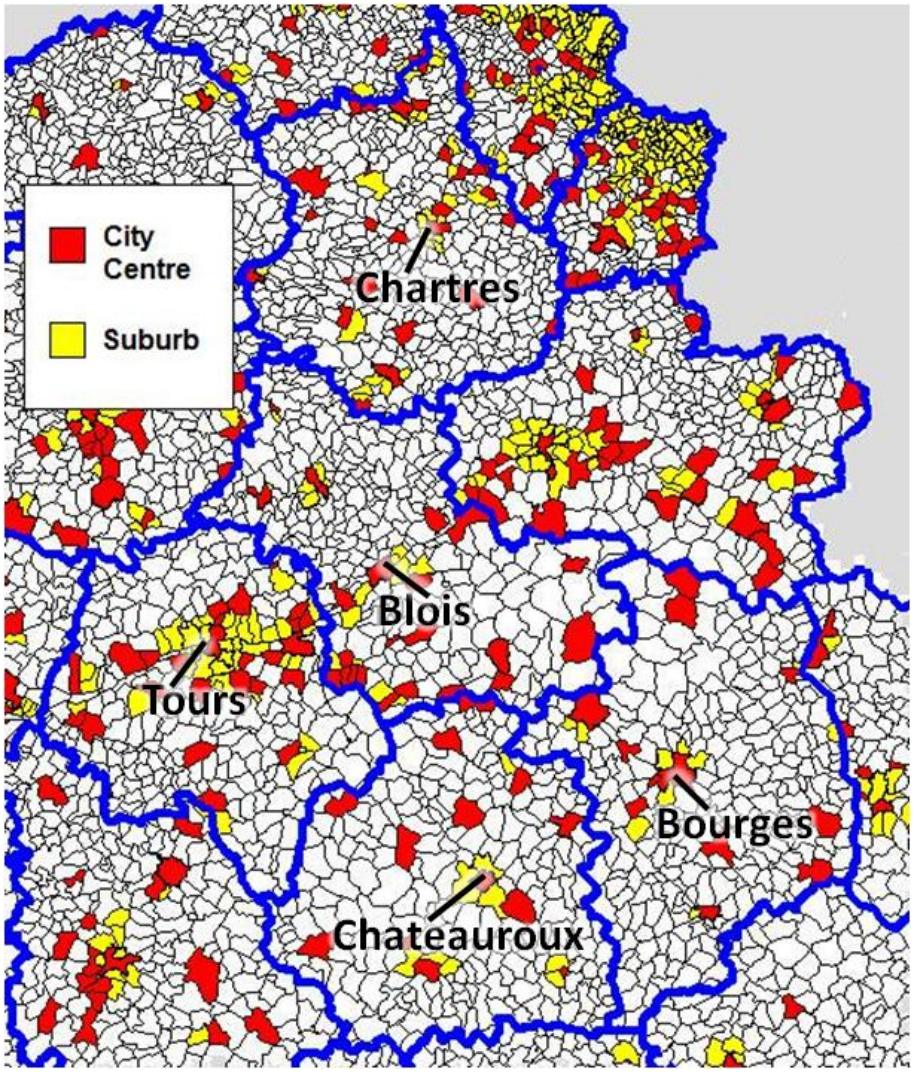

Figure 1.9: Centrality of communes in the Centre Region.

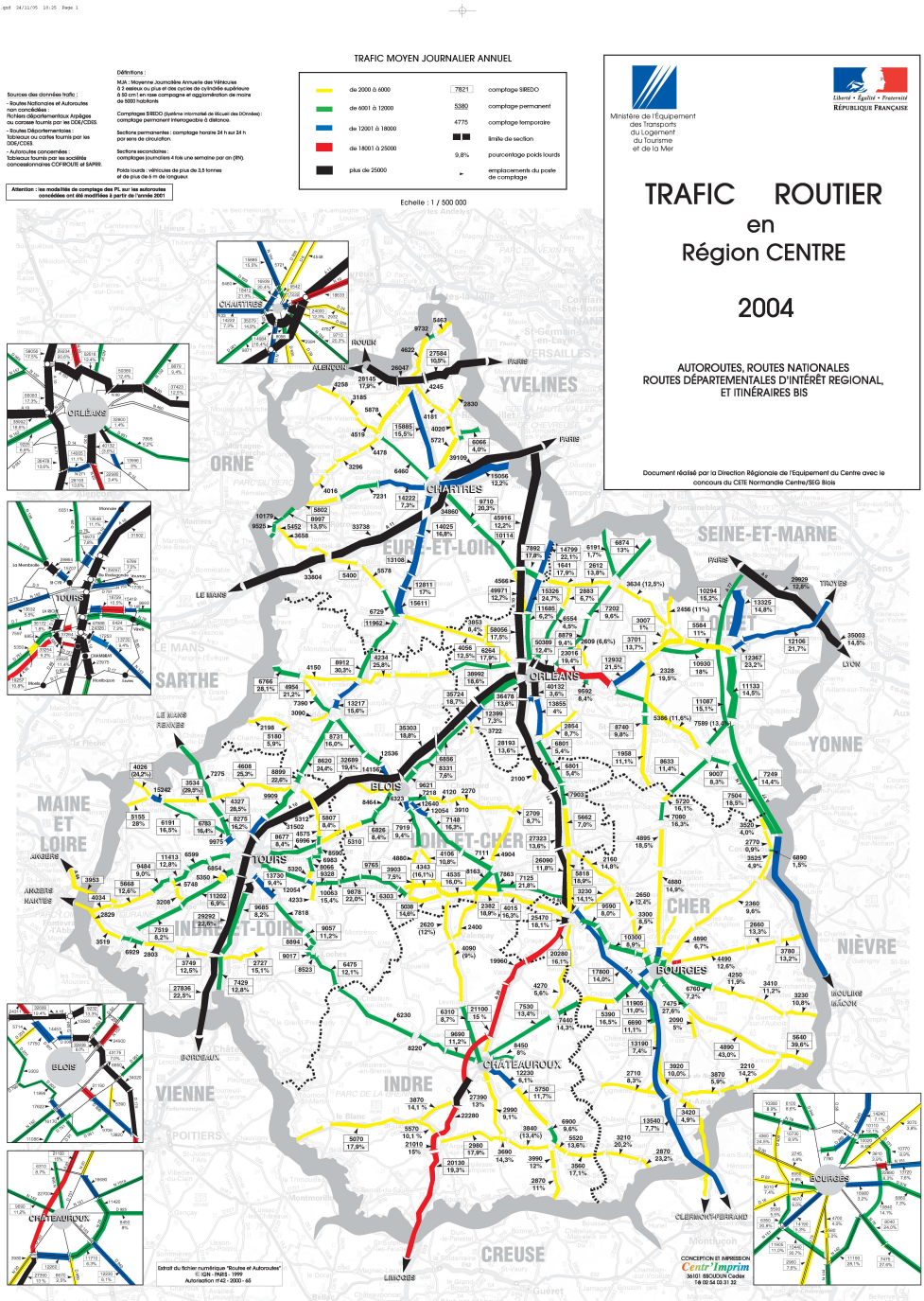

Figure 1.10: Vehicle traffic in the Centre Region of France.

of the study. As can be seen on the diagram, the roads in yellow had between 2000 and 6000 vehicles per day in 2006, the roads in green had 6001-12000, the roads in blue had 12001-18000, the roads in red had 18001-25000, and the roads in thick black had over 25000 vehicles per day. This serves as another way of showing the economic centralities of the cities of the region. The major highways in black are the main transportation axes that connect Paris with the South of France.

1.2 Discrete Choice Theory

The logit model is a model that is now widely used when modeling a choice that involves the selection of exactly one from a precisely defined finite set of alternatives. However, we begin by presenting the antecedent "gravitational" model of store attraction. According to this model, whose origins are attributed to the work of Reilly (1931 [39]), Luce (1959 [27]), and Huff (1963 [21]) the degree to which an individual is likely to choose a store can be quantified by a continuous variable, A_{nj}, that we call the "force of attraction" of the store, calculated using the following formula:

$$A_{nj} = \frac{s_j^\alpha}{d_{nj}^\beta}$$

where s_j is the size of the store j, which can be measured by retail space and d_{nj} is the distance between household n's domicile and store j. α and β are positive parameters of the function that are to be determined. The very intuitive implication of this is that the further a store is, the less a household will wish to visit it, and the larger a store is, the more a household will wish to visit it. For this reason, we suppose that the stores attract customers in the same way that celestial bodies attract satellites -in proportion to their mass and inversely proportional to a function of their distance. Originally conceived, this

model supposes that households choose the store that has exerted the greatest attraction upon it. However, Huff transforms this into a probabilistic model by applying Luce's Choice Axiom.

Luce's Choice Axiom (see Luce, 1977 [28]) states that if R is a subset of a set of alternatives S then the choice probabilities for the choice set R are equal to the choice probabilities of the choice set S conditional on a member of R being selected. That is,

$$P_R(a) = P_S(a|R) \forall a \in R, \forall R \subset S$$

since

$$\frac{P_R(a)}{P_R(b)} = \frac{P_S(a)P(R)}{P_S(b)P(R)} \forall a, b \in R, \forall R \subset S$$

this is equivalent to stating that there is a weight function w such that

$$\frac{P_R(a)}{P_R(b)} = \frac{w(a)}{w(b)} \forall a, b \in R, \forall R \subset S$$

Concretely, this means that the ratio of any two selection probabilities is proportional to the ratio of their respective weights w, or response strength, in Luce's terminology.

In his model of store choice, with C_n representing a choice set of stores for a given individual n, Huff assumes the Luce Choice Axiom, and takes as the response strength, the attractivity A_{nj} of store j in terms of "mass" and "distance", so that

$$\frac{P_{ni}}{P_{nj}} = \frac{A_{ni}}{A_{nj}} \ \forall i, j \in C_n$$

Because all alternatives in the choice set C_n are mutually exclusive and mutually

exhaustive,

$$\sum_{i \in C_n} P_{ni} = 1$$

and so

$$\sum_{i \in C_n} \frac{A_{ni}}{A_{nj}} P_{nj} = 1$$

From this, we can derive

$$P_{nj} = \frac{A_{nj}}{\sum_{i \in C_n} A_{ni}}$$

We note in passing that the attractivenesses of all the alternatives in C_n and their probabilities of selection must be nonzero, otherwise we will have a division by zero.

This means that the probability that a household will select a given store will be equal to the proportion of all the forces of attraction acting on the individual that is exerted by the given store. By allowing the introduction of other store characteristics besides size and distance, the Gravitational Model has been generalized into the Multiplicative Competitive Interaction Model (MCI) where if $X_{nj1}, \ldots, X_{njs}$ denote the s characteristics of a store j for household n, then the attractivity of the store will be

$$P_{nj} = \frac{X_{nj1}^{\beta_1} \cdots X_{njs}^{\beta_s}}{\sum_{k \in C_n} X_{nk1}^{\beta_1} \cdots X_{nks}^{\beta_s}}$$

where $\beta_1, \ldots, \beta_s$ are parameters to be determined.

The advantage of this model is that it is quite simple to calculate, and derives from intuitive assumptions about the relationship between a store's properties and a household's likelihood of selecting the store. We consider this a model that is of minimal simplicity, that can be used in order to reduce the costs of calculation. Logit models, in the next section, are an improved generalization of this model that are derived directly from assumptions about human behaviour.

A model which is very commonly used for a wide variety of discrete choice problems is the Conditional Logit Model. The reader is refered to Ben-Akiva and Lerman (1985 [6]) and Train (2003 [46]) for two references that provide a detailed presentation and derivation of this model. We assume that an individual n is faced with a known set of alternatives, that we denote C_n, of which it chooses exactly one. We then suppose that we can represent the criteria by which an individual evaluates any alternative j within C_n with a single numeric term called the utility (denoted by U_{nj}) and that the individual selects the alternative in the set for which this term is greatest. This means that if the utilities of a set of alternatives can be determined, then the choice of the individual can be predicted. Examples where the utilities are known might include the cases where someone chooses the cheapest of two sneakers of the same type, or the stock with the highest price to earnings ratio. These are cases where the value U of a decision is determined entirely by a known quantity, and the individual's decision can be known in advance. In the case of a household's choice of grocery store, the qualities considered in the household's decisions are unknown, and thus, we cannot predict which it will choose.

A Random Utility Model is applied to cases where uncertainty in utility determination precludes choice prediction. This model supposes that despite the utility being unknown, it follows a known random distribution. Thus, representing the utility by a random variable, we can calculate the probability that a utility associated with a given alternative will be greater than all other utilities corresponding to the other alternatives in the choice set. This probability is exactly the probability that the household will select the alternative in question. Thus,

$$P_{nj} \quad = \quad P(U_{nj} \geq U_{nk} \; \forall k \in C_n)$$

We suppose that the utility, U_{nj}, divides into a systematic component (V_{nj}), that we

calculate using known information, and a disturbance term (ε_{nj}), representing unknown factors used in an individual's decision-making, so that

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

1.2.1 The Logit Model

The Logit Model supposes that the disturbance terms are iid (independent and identically distributed), are independent of the systematic component, and follow the Gumbel Distribution within the family of Extreme Value Distributions. If X is a variable following the Gumbel Distribution, then the pdf of X will be

$$f(s) = e^{-x}e^{-e^{-x}}$$

and its cdf will be

$$F(x) = e^{-e^{-x}}$$

The important characteristic of this distribution is that the maximum of several independent extreme value distributed variables is also extreme-value distributed. But more importantly, if we suppose that the error terms of our random utilities are extreme-value distributed, then the probability of selecting a given alternative can be derived straightforwardly and expressed in closed form, as is seen below (also in Train, 2003 [46]):

$$\begin{aligned} P_{nj} &= P(U_{nk} \leq U_{nj}\ \forall k \in C_n) \\ &= P(\varepsilon_{nk} \leq V_{nj} - V_{nk} + \varepsilon_{nj}\ \forall k \in C_n) \\ &= \int_{-\infty}^{\infty} P(\varepsilon_{nk} \leq V_{nj} - V_{nk} + \varepsilon_{nj}\ \forall k \in C_n | \varepsilon_{nj}) f(\varepsilon_{nj}) d\varepsilon_{nj} \end{aligned}$$

And since the random variables are independent,

$$P_{nj} = \int_{-\infty}^{\infty} \prod_{k \in C_n} P(\varepsilon_{nk} \leq V_{nj} - V_{nk} + \varepsilon_{nj} | \varepsilon_{nj}) f(\varepsilon_{nj}) d\varepsilon_{nj}$$

Since

$$P(\varepsilon_{nk} \leq V_{nj} - V_{nk} + \varepsilon_{nj} | \varepsilon_{nj}) = \begin{cases} e^{-e^{-(V_{nj} - V_{nk} + \varepsilon_{nj})}} & , j \neq k \\ 1 = \left(\frac{e^{-e^{-(V_{nj} - V_{nk} + \varepsilon_{nj})}}}{e^{-e^{\varepsilon_{nj}}}} \right) & , j = k \end{cases}$$

$$\begin{aligned} P_{nj} &= \int_{-\infty}^{\infty} \prod_{k \in C_n} \frac{e^{-e^{-(V_{nj} - V_{nk} + \varepsilon_{nj})}}}{e^{-e^{-\varepsilon_{nj}}}} e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}} d\varepsilon_{nj} \\ &= \int_{-\infty}^{\infty} \exp \left\{ \sum_{k \in C_n} -e^{-(V_{nj} - V_{nk} + \varepsilon_{nj})} \right\} e^{-\varepsilon_{nj}} d\varepsilon_{nj} \end{aligned}$$

Replacing $e^{-\varepsilon_{nj}}$ with s (and $e^{-\varepsilon_{nj}} d\varepsilon$ with $-ds$) we obtain

$$\begin{aligned} P_{nj} &= \int_{\infty}^{0} -\exp \left\{ \sum_{k \in C_n} -s e^{-(V_{nj} - V_{nk})} \right\} ds \\ &= \left[\frac{\exp \left\{ \sum_{k \in C_n} -s e^{-(V_{nj} - V_{nk})} \right\}}{\sum_{k \in C_n} e^{-(V_{nj} - V_{nk})}} \right]_{\infty}^{0} \\ &= \frac{1}{\sum_{k \in C_n} e^{-(V_{nj} - V_{nk})}} - \frac{0}{\sum_{k \in C_n} e^{-(V_{nj} - V_{nk})}} \\ &= \frac{e^{V_{nj}}}{\sum_{k \in C_n} e^{V_{nk}}} \end{aligned}$$

The probability of household n selecting store j remains undetermined so long as we have not defined the term V_{nj}. The quality of our model will depend upon the degree to which we are able to define the systematic component of the utility term in a way that accurately reflects a household's behaviour pattern. This term represents what we know

about a given household.

In order to define V_{nj}, we make use of the Conditional Logit Model, and suppose that it is determined by a linear function of several explanatory variables reflecting the characteristics of the alternative for the individual in question so that:

$$V_{nj} = \beta X_{nj}$$

where X_{nj} is a vector of several explanatory variables, and β is a vector of coefficients of the explanatory variables. An alternative to this approach would be to use the Multinomial Logit Model, in which V_{nj} would be defined

$$V_{nj} = \beta_j X_n$$

where X_n is a vector of explanatory variables that depend on the individual n, and are independent of the alternative j, and β_j is a vector of parameters that determines the attractiveness of alternative j for n. In both cases, we are assuming a linear-in-parameters model, which is standard in the literature. However, the utility could be defined using more complex functions of the explanatory variables. Abe (1999 [1]), for example, proposes to replace a linear combination of explanatory variables with a sum of one-dimensional nonparametric functions of the explanatory variables.

In general, the Multinomial Logit Model is used in the case where there is data on the characteristics of decision-makers, and every individual is faced with the same alternatives in their choice sets. This is used typically in modelling choices of brands amongst consumers. The Conditional Logit Model gets its name from the fact that it is typically used where we wish to model a choice conditional on the characteristics of the individual making the decision. That is, it is based on the characteristics of the

alternatives, and not of the individual. In our case, we believe that a Conditional Logit Model is more appropriate for the prediction of store choices, since the set of stores varies by individual, and the most important factors that we include in our model are also store characteristics such as distance, that also vary greatly from one individual to another. [5]

The X_{nj} vector includes such variables as distance from the household's home, store size, etc. The right choice of variables to include in this vector is of primordial importance, as is the determination of the values contained in β. The validity of the assumption that the disturbance terms in the utilities are iid depends upon our ability to choose a set of variables that account for enough of all the possible factors involved in the choice of alternatives made by the individual without introducing irrelevant information.

The β vector remains unknown, so we generate estimates of its value through the use of maximum likelihood estimation. This technique relies on a sample S of individuals for which the choice of large-surface store is known. The variable z_{nj} will be one if n is in S and j is the alternative selected by n, and zero, otherwise.

If the predicted probability that n selects j conditional on the value of β is

$$P_{nj}(\beta) = \frac{e^{\beta X_{nj}}}{\sum_{k \in C_n} e^{\beta X_{nk}}}$$

then the maximum likelihood estimate of β will be the value of β that maximizes the

[5] In SAS, version 9, Multinomial Logit estimation can be done using the CATMOD Procedure and Conditional Logit estimation can be done using the MDC ("Multinomial Discrete Choice") Procedure. Data sets used as input for the CATMOD procedure are organized by individual, each observation containing one variable that indicates the value of the observed response for the individual. For the MDC Procedure, each observation corresponds to an alternative presented to an individual, and a binary variable indicates whether the alternative is selected by the individual or not. With some manipulation, MDC can be used to estimate a Multinomial Logit model, although this requires the definition of a much more cumbersome data set than is the case for CATMOD. However, there are many problems, such as the one treated in this thesis, upon which the MDC Procedure in SAS can be applied, but not the CATMOD Procedure.

log of the likelihood of the sample responses

$$\begin{aligned} \mathrm{LL}(\beta|S) &= \sum_{n\in S}\sum_{j\in C_n} z_{nj} \ln P_{nj}(\beta) \\ &= \sum_{n\in S}\sum_{j\in C_n} z_{nj} \ln\left(\frac{e^{\beta X_{nj}}}{\sum_{k\in C_n} e^{\beta X_{nk}}}\right) \end{aligned}$$

The reader can consult Appendices 4.4 and 4.4 for more details on the basic derivations of the estimation formulae for the Conditional Logit models. This expression is expressed as being conditional of the attributes of the sample S.

An advantage of the conditional logit model is that it has become classic. It is well-known and well-understood, its derivations and justifications are intuitive, and it is mathematically simple to manipulate, making it a useful model to use in BVA's product development.

An added advantage of this model is that the familiar gravitational of store choice is a special case. With $\tilde{\beta} = (\alpha, -\beta)$ and $x_{nj} = \begin{pmatrix} \log(s_j) \\ \log(d_{nj}) \end{pmatrix}$, we obtain

$$\exp(\tilde{\beta} x_{nj}) = \exp(\alpha \log(s_j) - \beta \log(d_{nj})) = \frac{s_j^{\alpha}}{d_{nj}^{\beta}}$$

which is the formula for the force of attraction of store j for household n in the gravitational model.

The problem with this model is that it has simplifying assumptions in its derivation that may not be verified. The most widely-discussed of these assumptions is the assumption of independence of irrelevant alternatives (IIA). This assumption, which is in fact equivalent to Luce's Rule, states that if i and j are two alternatives in the choice set C,

and C' is a choice set containing C, then:

$$\frac{P(i|C)}{P(j|C)} = \frac{P(i|C')}{P(j|C')}$$

where $P(i|C)$ is the probability of selecting i from the choice set C. [6] What this entails is that the ratio of the probabilities of selection of two alternatives is unaffected by the addition of new alternatives to the choice set nor any change in the attractiveness of another alternative.

Let us look at an example. Suppose a household is twice as likely to shop in store A than in store B. Now suppose that there is a third store, C, that its owners decide to expand and renovate, with the effect that the likelihood that our household shops in the store increases by threefold. Then according to IIA, the household will be less likely to select either store A, or store B, but will still be twice as likely to select store A as store B. This may seem reasonable, since the intrinsic attractivenesses of store A and B has not changed, and therefore we can say that the relative attractiveness of A with respect to B ought to remain the same. However, this assumption would be false if there existed spatial correlations between store preferences, meaning that a store's utility would be affected by the utility of a nearby store. An example of this could be where the renovations of store C attract customers from far away, and thus increase the utility of all stores close to it by exposing them to customers who would not normally enter into the town in which they are located. This would mean that if store A is close to

[6] Since

$$\begin{aligned} \frac{P(i|C)}{P(j|C)} &= \frac{\frac{e^{V_{ni}}}{\sum_{k\in C} e^{V_{nk}}}}{\frac{e^{V_{nj}}}{\sum_{l\in C} e^{V_{nl}}}} \\ &= \frac{e^{V_{ni}}}{e^{V_{nj}}} \end{aligned}$$

for any choice set C

store C, but store B is not, then the increase in utility of store C will cause a greater increase in the utility of store A than store B, violating IIA. Another case where IIA is not verified may be when there is a correlation between similar types of stores. For example, if a store of a high-valued goods is introduced, there may be more customers taken from other stores of high-value goods, than other types of stores.

There are variations of the basic Conditional Logit model that have been developed in order to create models that can account for the fact that error terms cannot be assumed to be independent and thus could deal with these problems. These are also found in Train(2003 [46]) and Ben-Akiva and Lerman (1985 [6]).

1.2.2 Weighting for sampling

In finding values of the β vector that are best adapted to a sample of individuals, we have not taken into account the fact that our sample is not representative of our population. Effectively, this means that maximum likelihood estimation maximizes the model likelihood of observed responses within the sample, but we have not shown that it maximizes the model likelihood of responses within the population.

We have at our disposal a set of weights that allow us to adjust the population profiles of the sample so that they match those recorded in census data on the survey region. We have considered using these weights in order to adjust our maximum likelihood estimate to make it more representative of the population. One way we could do this would be to treat our sample as if it had been produced through random sampling and the weights corresponded to the inverses of the sampling probabilities of the individuals.

Ben-Akiva and Lerman (1985 [6]) have shown that if exogenous random sampling is used to construct the sample, that the maximum likelihood of the alternative choices by sampled individuals do not depend on their probabilities of inclusions in the sample, and

so no use need be made of individual weights.

The log-likelihood estimate in Equation 1.1 can actually be viewed as an expression that is conditional on the distribution of attributes in the sample S. Thus, we can rewrite $P_{nj}(\beta)$, the function expressing the probability that n chooses j conditional on the model parameter β, as $P(j|X_n, \beta)$, the probability that alternative j is selected conditional on the model parameter β and X_n, the attributes of the choice set C_n of individual n. So this means that Equation 1.1 becomes

$$\mathrm{LL}(\beta|S) = \sum_{n \in S} \sum_{j \in C_n} z_{nj} \ln P(j|X_n, \beta) \tag{1.1}$$

An alternative log-likelihood expression that is not conditional on the sample S would be the log of the joint likelihood of alternatives selected in sample S and the distribution of attributes in S. This would be written

$$P(j, X_n|\beta) = P(j|X_n, \beta)\pi(X_n)$$

where $\pi(X_n)$ represents the distribution of the attributes of X_n. In exogenous random sampling, this value corresponds to the sampling probability of n, and does not depend upon the parameter β. This gives us an unconditional log likelihood expresson

$$\begin{aligned} \mathrm{LL}(\beta, S) &= \sum_{n \in S} \sum_{j \in C_n} z_{nj} \ln P(j, X_n|\beta) \\ &= \sum_{n \in S} \sum_{j \in C_n} z_{nj} \ln P(j|X_n\beta) + \sum_{n \in S} \pi(X_n) \end{aligned}$$

Since $\pi(X_n)$ is constant with respect to β, finding the parameters that maximize $\mathrm{LL}(\beta, S)$ is equivalent to finding the parameters that maximize $\mathrm{LL}(\beta|S)$. Thus, in ex-

ogenous random sampling, maximum likelihood estimates do not take into account the sampling strategy used, and therefore no weighting adjustment is necessary.

Where the use of weights in maximum likelihood estimates could be justified would be where we wish to maximize the model likelihood of all store choices in the general population U, as in the following formula

$$\mathrm{LL}(\beta|U) \quad = \quad \sum_{n\in U}\sum_{j\in C_n} z_{nj} \ln P_{nj}(\beta)$$

Since z_{nj} is unknown, we would have to approximate this expression. According to basic sampling theory (a brief discussion of this is in Appendix 3, if we consider $\sum_{j\in C_n} z_{nj} \ln P_{nj}$ for all values of n to be independent and identically distributed random variables, then an unbiased estimator of $\mathrm{LL}(\beta|U)$ would be

$$\hat{\mathrm{LL}}(\beta|U) \quad = \quad \sum_{n\in S}\sum_{j\in C_n} w_n z_{nj} \ln P_{nj}(\beta)$$

where S is a random sample drawn from U, and w_n will be the inverse of the probability that n is drawn from U. When we use our linear-in-parameters Conditional Logit Model, this becomes

$$\hat{\mathrm{LL}}(\beta|U) \quad = \quad \sum_{n\in S}\sum_{j\in C_n} z_{nj} \frac{e^{\beta X_{nj} + \ln w_n}}{\sum_{k\in C_n} e^{\beta X_{nk}}}$$

The maximum likelihood estimate of a parameter is also the parameter that sets the estimated covariance between residual terms and explanatory variables to zero. In Appendix 4, we show how Equation 1.2 also arises by adjusting this covariance term so that the profiles of the explanatory variables match those within the general population.

We ought to take care before dismissing the use of weights in Conditional Logit estimation, however. We have shown that the use of weights does not affect the maximum likelihood estimates of the model parameters, but this does not mean that the weights do not affect the maximum likelihood estimates of the variances of the parameter estimates, and these are later used in t-tests that we use to determine which parameter estimates are to be used in our model.

In general, the arguments between weighting and not weighting an estimator come down to the way sampling is viewed. Those favouring weighting would like to take into account the way in which a sample was constructed in order to create the model. Those favouring no weighting believe that maximum likelihood involves a test of the model against a sample and we should not make assumptions about how the rest of the population behaves. They would also point out that with severely unbalanced weights, the maximum likelihood estimations could be dominated by a few individuals, making it more vulnerable to measurement error.

In the future, in the interests of completeness, a comparison of our modelling results can to be done both with and without using the sample weights, however, we have decided not to weight our estimations. It is not the likelihood of the alternatives selected in the population that concerns us, since we are concerned with our model's predictive power, when applied to any sample of individuals drawn from anywhere in France, and not its fit to the data upon which it is estimated. What we are concerned about is looking at evidence of possible causal relations between explanatory variables and observed behaviour and applying these to any population we wish. Even if there were a strong case to be made for weighting our estimator, the fact that our survey sample was collected through quota, rather than random sampling means that the assumptions underlying our weighting technique are greatly weakened.

1.3 Summary

The BVA survey of shopping flux in the Centre Region was done on a sample of households selected through quota sampling, with quotas defined for survey sectors and household types. Survey sectors were defined by BVA by aggregating small groups of Census districts called "IRIS". Once a survey sample was determined, chosen households would be asked in which three large-surface store choices they visited the most often, the second most often, and the third most often. The responses would then be matched to a data set of all large-surface stores within the survey region. These large-surface stores were divided into three categories, supermarkets, hypermarkets, and hard discount stores. Hypermarkets were further divided into two groups, small hypermarkets, and large hypermarkets. Auxiliary information, including the socio-demographic characteristics of each member of each household were recorded by BVA, along with the known characteristics of all large-surface stores in the area. More information could be obtained from public data such as an INSEE data set that contains a sample of one in 20 1999 census respondents. Importantly, the geographic co-ordinates of both household domiciles, and store locations were recorded, allowing us to calculate the distance between a household's home and every store in the neighbourhood and determine which store was the closest. Unfortunately, there were limitations in the accuracy of these co-ordinates, since they were assigned by "IRIS".

We found in initial studies of our data, that households were very influenced by the distance of a store from its homes. Over half of all households in the sample selected the large-surface store that was closest to its home within its category. This meant that one of the biggest determinates of store choice would be the rank of the store's distance from the household's home with respect to other stores.

We decided to use Conditional Logit in order to predict each household's store choice.

This model assumes that each household assigns a utility to each alternative considered in its choice set and that this utility is composed of a systematic component and an error term that is a random value following the Gumbel Distribution. This model specification allows us to express the model predicted probability in a closed form. We assume that the systematic component of the utility is determined by a set of explanatory variables taken from our survey data. The parameters of our model are estimated using maximum likelihood estimation. In order to use this type of model, we need to assume the independance of irrelevant alternatives property, which assumes that the introduction of new alternatives in the choice set, or the elimination of old alternatives will cause the probabilities of all remaining alternatives to change proportionately. This is a condition we believe to be verified in general in large-surface store choice. We considered weighting our sample in order to take into account the effect of sampling on the household, however, this idea was abandoned since we were doing exogeneous sampling, which meant that our maximum likelihood estimator was unaffected by sampling strategy.

Chapter 2

Choice set definition and criteria of model evaluation

In Chapter 1, we introduced the Conditional Logit model in theory. However, in this chapter, we look at some theoretical issues that arise when we apply this model to our data. A Conditional Logit model presupposes a choice set of exhaustive and mutually exclusive alternatives that is exactly defined. There are two issues that we look at in detail: the definition of a choice set, and the evaluation of the quality of our model prediction. When we wish to apply this to our data, we must then determine what alternatives correspond to our model of store choice. If we wish to predict which large-surface store a household will choose, we must know in advance from which alternatives the household will choose from. In Section 2.1, we describe how this is a problem in our case, since a household can in theory select any of well over a thousand different large-surface stores, but will only really consider a subset of large-surface stores. We therefore make a distinction between a universal choice set of all the alternatives for which a store choice cannot be ruled out, and the consideration set, which is the unknown set of stores from which the household actually chooses from. The universal set is known to

the statistician, but unfortunately, the consideration set is not. We therefore introduce several ways in which our model can be changed in order to take this fact into account. In Section 2.2, we look at how to evaluate our model once we apply it to our data. There are a few measures that we can use. One would be the McFadden Pseudo-R-Squared, based on goodness of fit, which calculates how well the model's predicted probabilities fit with the data set upon which it is based. Since we wish to use our model to calculate the accuracy of predictions made on other data sets, rather than how well it corresponds to the data set upon which it is based, we prefer using cross-validation of our model predictions by using one part of our survey data to estimate the parameters of our Conditional Logit model that we will use to generate predictions on the other part of the data set, which we can then compare with the observed values for these individuals. In this section, we describe the "WD" statistic that we use to measure the difference between the forecast number of households from one geographic area visiting a given large-surface stores and the observed number.

2.1 The problem of large choice sets

A difficulty with logit and gravitational models is that they can only be used to model decisions for which we have precisely defined a choice set of mutually exclusive and exhaustive alternatives whose properties are known well enough that a utility can be calculated for each one. Unfortunately, a set of grocery stores that include all the possible choices made by households for their food purchases will have to include the entire set of all stores recorded. There are a few reasons why this ought not be done. First of all, it is unreasonable to believe that a household considers every single one of the 1600 large-surface stores recorded within our survey of household store choice in making its decision. The household is surely not even aware of the existence of many of these stores.

Thus, our model has unrealistic assumptions about human behaviour. Second, although it is possible for the household to select any large-surface store, many of the probabilities of selection for these stores (such as those that are very far away from the household's home) are extremely low. Due to the nature of maximum likelihood estimation, the lower the probability that individuals select the alternatives they have been observed to select, the greater the weights these individuals will have in model estimation relative to other individuals.[1]

However, the biggest problem with large choice set sizes is simply that the computational burden is too great for them to be feasible, and so we are forced to cut down the number of alternatives that we specify for each decision-maker.

In this chapter, we shall discuss different ways in which we can redefine the choice sets of each individual, or how to reduce the computational burden of large choice sets. However, once we have done this, we will need to choose between the different approaches we shall use. Doing this is not easy, since more traditional methods of evaluating models such as the likelihood ratio, or the McFadden Pseudo R^2, or Brier Score are not appropriate for evaluating different choice set definitions for the same model. This, unfortunately, means that the question of choice set definitions we use is inextricably linked with the development of methods that we shall use to evaluate these choice sets.

We can distinguish between a universal choice set, that is, a set containing all alternatives that are theoretically possible for the choices of stores associated with every

[1]This can be seen quickly with the formula for maximum likelihood estimation from Equation 1.1:

$$\mathrm{LL} = \sum_{n=1}^{N} \sum_{k \in C_n} z_{nk} \ln(P_{nk})$$

As any value P_{nk} goes to zero, $\ln(P_{nk})$ goes to negative infinity, and small changes in the values of P_{nk} that are very small will have a much greater effect on LL than for values of P_n that are not. Thus, if we include alternatives with very low probabilities of selection within the choice set, we risk allowing outliers or erroneous observations (that is, erroneous values of z that fall on alternatives with low probabilities) to distort the sample.

individual (in our case, every single large-surface store recorded in our survey) and the consideration set, that is, the subset of the universal choice set containing the alternatives we believe the household considers in making its choice. The fact that the consideration set of large-surface stores is unknown is our challenge.

There have been several methods proposed for estimation when either the consideration set for each individual is unknown, or the universal choice set is too large. We look at three of these: using random consideration sets, aggregating alternatives, and sampling alternatives. The first technique is not possible in our case, so we only calculate the last two methods. Most of this section will deal with sampled alternatives, due to the greater technical detail involved in its explanation.

2.1.1 Using random consideration sets

The use of a randomly-defined consideration set is done in order to take into account the fact that an individual will not consider all possible alternatives in making its decision. Thus, according to our model, each individual n will select one alternative from a consideration set C_n following the Logit Model, so that

$$P_{nj} = \frac{e^{V_{nj}}}{\sum_{k \in C_n} e^{V_{nk}}}$$

However, although the universal set of all possible alternatives U_n is known, we do not know the consideration set C_n. A simple solution would be to assign a probability P_C to each possible choice set C which is a subset of U_n so that

$$\sum_{C \subset U_n} P_C = 1$$

The probability of a household choosing a given alternative will thus be determined by

Bayes' theorem.

$$P_{nj} = \sum_{C \subset U_n} P_C P_{nj|C}$$

with

$$P_{nj|C} = \frac{\exp(\beta x_{nj})}{\sum_{k \in C} \exp(\beta x_{nk})}$$

which is a Logit Model conditional on the consideration set defined. Provided that the probabilities P_C are well-developed, this two-stage model may serve as an improvement to a model in which we assume that all individuals base their decision on the assigning of utilities to all possible alternatives. An early developement of this model can be found in Manski (1977 [29]). Swait and Ben-Akiva (1987 [44]) developed a random constraint model in which alternatives are excluded from the choice set if their utility falls below a threshold level. This threshold level is a model parameter that is estimated along with the parameters of the conditional logit model. A recent application of this last model is found in a paper by Basar and Bhat (2004 [5]) who use this random constraint model, but allow the inclusion threshold to vary by individual.

Basar and Bhat state that their model is superior to the model that assumes that individuals consider every single alternative in the universal choice set, since it adjusts for the fact that individuals only take into account a subset of the universal set of alternatives in making their decision. It would seem logical to apply this type of model to the choice of supermarkets in France, since we are dealing with choice sets that are obviously too large to correspond to actual consideration sets. Unfortunately, this model addresses the problem of unrealistic consideration sets, but the greater model specification comes at the cost of massively increased computational burden. As the size of U_n increases, the power set $\mathcal{P}(U_n)$ of all possible subsets of U_n rapidly becomes too large for calculations to be possible, meaning that it is obviously inappropriate for universal choice sets containing hundreds of supermarkets. The reason this is applicable in the Basar and Bhat paper is

that the problem involves a universal choice set consisting of three airports in the San Francisco Bay Area, and so the number of possible consideration sets is only $2^3 = 8$.

2.1.2 Aggregating choices

Instead of randomly selecting alternatives in a choice set, a way of reducing the computational burden of Logit Model estimation is to redefine the choice set under consideration so that it contains a workable number of alternatives for estimation while allowing the model predictions to remain sufficiently informative.

This technique is described by Ben-Akiva and Lerman [6], pp. 253-261. Using this approach, we assume that the Conditional Logit model applies to a universal set U_n of alternatives that are termed "elemental" alternatives. Thus,

$$P_{nj} = \frac{e^{V_{nj}}}{\sum_{k \in U_n} e^{V_{nk}}}$$

Suppose that o is a subset of U_n. We define P_{no} as the probability that individual n selects an alternative that is contained in o. Let ι be equal to the subset of all alternatives in U_n but not in o.

$$\begin{aligned} P_{no} &= \sum_{j \in o} P_{nj} \\ &= \frac{\sum_{j \in o} e^{V_{nj}}}{\sum_{k \in U_n} e^{V_{nk}}} \\ &= \frac{e^{V_{no}}}{e^{V_{no}} + \sum_{k \in \iota} e^{V_{nk}}} \end{aligned}$$

where

$$V_{no} = \ln\left(\sum_{j\in o} e^{V_{nj}}\right)$$

When this formula is either too long to calculate due to o containing a large number of elemental alternatives, or impossible, if the data used to generate the systematic components of utility is not available, the technique of aggregating alternatives consists in replacing V_{no} with an approximation, $\tilde{V}_{no}$. We may suppose, for example, that

$$\tilde{V}_{no} = \alpha z_{no}$$

where z_{no} could consist of

$$z_{no} = \frac{1}{|o|}\sum_{j\in o} x_{nj}$$

the average value of the vector of explanatory variables on the alternatives within the choice set o, or it could consist of entirely other variables that are good proxies for the utility of the alternatives in o.

We have used this technique in a model of shopping behaviour we developed earlier that reduces choices of stores to simply the selection of a general type of store (supermarket, hypermarket, or hard discount), and the option of selecting no stores, in which we did not enter variables reflecting the characteristics of the stores themselves. We found that this particular model performs very poorly since it does not take into account the distance between homes and individual stores. This technique is too restrictive, causing us to lose too much valuable information. A better approach, favoured by Howard Smith (2004 [41]) is to take advantage of the fact that households are overwhelmingly

more likely to select stores that are near the household's domicile than far. He selects the 30 closest supermarkets to each household as his choice set, and then includes a thirty-first alternative that corresponds to all other store choices. Indeed, he has created a new alternative through the aggregation of all far stores, yet his choice set still retains important information on those stores that represent an overwhelming number of store choices.

This is the solution that seems most practical to us, since it is the easiest method to estimate, and the choice sets probably correspond best with the likely consideration sets of each individual. However, we must adjust make some basic adjustments to this technique if we are to apply it to large-surface stores in France. First of all, if we take only the closest stores to each household, stores that might have an exceptionally large attraction risk not to be included in the choice sets of households liable to consider them. If someone lives in the interiour of a very large city, there could very well be more than thirty supermarkets between the household's home and the nearest "big-box" store the household will visit. Secondly, since we are working on a model intended to predict the actual store a household visits, if a household is predicted to select the outside option, then we do not have a prediction of the actual store a household is going to select.

2.1.3 Sampling alternatives

The maximum likelihood estimates of our conditional logit model, if they can't be calculated using the actual maximum likelihood, can still be estimated with a modified maximum likelihood formula in which the choice sets of each individual are replaced with a random sample of the alternatives found in each choice set, and through the inclusion of adjustment terms. This is the conclusion of a paper by McFadden (1978 [34]) in which he shows that assuming the independence of irrelevant alternatives property

holds, and given certain conditions, the parameter estimates calculated over these sampled alternatives will be consistent with the true model parameters. This means that we can still estimate a conditional logit model even in cases where the choice sets of each individual are too large, or even unknown.

An application of this can be found in a paper by Train, McFadden, and Ben-Akiva (1987 [47]) who face a universal choice set that is infinite, and consideration sets that are unknowable. This paper describes a model of the selection of service plans by telephone company clients. Since the cost of each phone service depends on phone usage, the authors consider each client as selecting a combination of a phone service plan and a phone usage pattern. Since the set of all possible plans and usage patterns for each individual cannot be defined, a Conditional Logit expression cannot be calculated over a universal choice set. The solution the authors find is to create their choice sets by drawing random samples from a prior distribution of plans and usage patterns. The parameters of the model are estimated using Logit estimation on these choice sets. We now take the time to explain sampling alternatives in detail. We begin by recalling that the probability that household n selects store j, given the parameters of the model β will be:

$$P(j|n,\beta) = \frac{e^{V_{nj}(\beta)}}{\sum_{k\in U} e^{V_{nk}(\beta)}}$$

where U is the universal choice set. A maximum likelihood estimate of β over a sample of N individuals is the value of β that maximizes the expression

$$\mathrm{L}(\beta,N) = \prod_{n\in N}\prod_{j\in U}\left(\frac{e^{V_{nj}(\beta)}}{\sum_{k\in U} e^{V_{nk}(\beta)}}\right)^{z_{nj}}$$

where z_{nj} is one when n selects j, and zero, otherwise. Now, suppose that we construct

a choice set C of a restricted number of alternatives that includes the alternative selected, and a set of alternatives drawn at random from the set U. $P(j|C,n)$ will be the probability that household n selects alternative j given that the restricted choice set we generated is C. Similarly, we can define $P(C|n,j)$ as $P(C|j \in C, C \subset U, n)$. This probability is zero if j is not in C. Then we find that

$$\begin{aligned}
P(j|C,n) &= \frac{P(j,C|n)}{P(C|n)} \\
&= \frac{P(j,C|n)}{\sum_{k\in C} P(C,k|n)} \\
&= \frac{P(C|j,n)P(j|n)}{\sum_{k\in C} P(C|k,n)P(k|n)} \\
&= \frac{P(C|j,n)\frac{e^{V_{nj}}}{\sum_{l\in U} e^{V_{nl}}}}{\sum_{k\in C} P(C|k,n)\frac{e^{V_{nk}}}{\sum_{l\in U} e^{V_{nl}}}} \\
&= \frac{P(C|j,n)e^{V_{nj}}}{\sum_{k\in C} P(C|k,n)e^{V_{nk}}} \\
&= \frac{e^{V_{nj}+\ln P(C|j,n)}}{\sum_{k\in C} e^{V_{nk}+\ln P(C|k,n)}}
\end{aligned} \tag{2.1}$$

$P(C|j,n)$ is the probability that we will generate the choice set C for household n if j is chosen by n. $P(j|n)$ is the probability of household n selecting alternative j according to our model's original assumptions, and $P(j,C|n)$ is the joint probability of household n selecting alternative j and the set C being constructed for n.

We can see that this is an ordinary Conditional Logit Expression, only there are weights that take into account the discrepancy between the probability of an alternative's inclusion in the restricted choice set conditional on the alternative being selected, and the probability conditional on the alternative not being selected.

Let us look at a small example in order to get a better intuition. Suppose we take as an alternative "Supermarket X", which is in a set U of 100 alternatives. Suppose

that there are 1000 alternatives in the sample, meaning that there are 1000 choice sets defined, one for each individual. If "X" is chosen by 10 percent of households, then 10 percent of the choice sets in this model containing Supermarket X correspond to a choice of Supermarket X. Suppose that for each household n, we construct a reduced choice set, C_n, constructed by including the selected alternative and then adding a set of 10 alternatives drawn at random from U. Suppose that if X is not chosen by n, that X has a one in ten chance of being included in C_n. The expected number of households having choice sets containing X will then be the number of households selecting X ($0.10 \times 1000 = 100$) plus the expected number of the remaining households for whom X was drawn as one of the nonchosen alternatives in its choice set ($0.10 \times 900 = 90$), making a total of 190. This means that the percent of households whose choice set contains X that choose X will now be expected to be $100/190 = 52.6\%$. Therefore, by restricting our choice set, we inflate the apparent probability of selecting X from 0.10 to 0.526. The lower the probability that an alternative will be in the restricted choice set if it is not chosen, the greater this "inflation" will be. Suppose that another ten percent of households choose Y as an alternative, but that Y has a one in two chance of being selected as a nonchosen alternative if it is not chosen by a given household. This means that the number of households having choice sets containing Y will be equal to the number selecting Y ($0.10 \times 1000 = 100$) plus the expected number of the remaining households for whom Y was drawn as one of the nonchosen alternatives in its choice set ($0.50 \times 1000 = 500$) making a total of 600. Thus, the expected percent of households having Y in their restricted choice set choosing Y is $100/600 = 0.167$. Thus, although X and Y have the same probability of being selected, by restricting our choice sets using unequal sampling probabilities, the resulting apparent probabilities of selecting X and Y are far different, and will cause us to overestimate the utility of X with respect to Y in cases where X and Y are both

included in the same restricted choice set.

In order to take this effect into account, if X and Y both belong to C_n, the utilities of X and Y are handicapped through the addition of a negative term corresponding to the log of the probability that C_n would have been generated had the alternative in question been the alternative selected by the household. Indeed, since Y has a probability of being included when not chosen by the household that is five times greater than X's probability, the probability of C_n arising when Y is the chosen alternative ($P(C|Y,n)$) is greater than the probability of C_n arising when X is chosen ($P(C|X,n)$). This means that we subtract a greater term from V_{nY} than from V_{nX} when calculating the utilities of X and Y in C_n. We need to show now that this use of the weighting of drawn alternatives effectively eliminates parameter bias.

There are two consequences of 2.1. First of all, $P(C|j,n)$ must be strictly greater than zero, otherwise our expression will include logs of zeros and be undefined. This leads to the following definition:

Property 1 (Positive Conditioning) *For any individual n, if* $j \in C \subset U$ *and* $P(C|n,i) > 0$*, then* $P(C|n,j) > 0$

Concretely, this means that provided that it is logically possible for a given choice set to be assigned, then it is logically possible for this choice set to have been assigned to n no matter which alternative within this choice set was actually observed for individual n.

The second consequence of 2.1 is that if $\ln P(C|j,n)$ remains constant for all j in C, then these terms cancel out of the expression, and

$$P(j|C,n) = \frac{e^{V_{nj}}}{\sum_{k \in C} e^{V_{nk}}}$$

meaning that we can use the same Logit expression on the reduced choice set C that we did on U without needing to make an adjustment for our sampling strategy. This brings us to the following definition:

Property 2 (Uniform Conditioning) *If $i, j \in C \subset U$, then $P(C|n, j) = P(C|n, i)$.*

This is clearly a special case of the Positive Conditioning Property. Concretely, this means that no matter what choice of alternative was observed for individual n, the probability of assigning the choice set in question remains the same.

Whenever we develop a technique for sampling alternatives, we need to check these two properties. If a sampling strategy satisfies the Positive Conditioning Property, then it is possible to use this strategy in a modified likelihood formula in order to produce estimators that are consistent with the true model parameters. If a sampling strategy also satisfies the Uniform Conditioning Property, then we do not have to calculate the associated sampling probabilities in order to create a modified likelihood formula that produces consistent estimators.

Consistency of estimators

The proof is taken from pages 545-546 of an article written by McFadden (McFadden, 1978 [34]). We follow this proof in slightly more detail here. z_{nj} is the familiar indicator variable that is one when n chooses j, and zero otherwise.

Theorem 1 (Consistency of Estimations on Sampled Alternatives) *If for all n and for all k in C_n, $P(C_n|k)$ satisfies the positive conditioning property, and our model satisfies the Independence of Irrelevant Alternatives Property (IIA) (or Luce's Rule) as in the case of a Conditional Logit model, then the values of β that maximize the following*

modified log likelihood function

$$LM_N(\beta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{j\in C_n} z_{nj}\ln\left(\frac{e^{(V_{nj}(\beta)+\ln P(C_n|j))}}{\sum_{k\in C_n} e^{V_{nk}(\beta)+\ln P(C_n|k)}}\right)$$

are consistent with the values of β^* that maximize the original choice set of possible alternatives:

$$L_N = \frac{1}{N}\sum_{n=1}^{N}\sum_{j\in U} z_{nj}\ln\left(\frac{e^{V_{nj}(\beta^*)}}{\sum_{k\in U} e^{V_{nk}(\beta^*)}}\right)$$

Proof:

With the uniform conditioning property, $P(C_n|j) = P(C_n|k)\ \forall\ j,k \in C$, and so we have:

$$LM_N(\beta) = \frac{1}{N}\sum_{n\in N}\sum_{j\in C_n} z_{nj}\ln\left(\frac{e^{V_{nj}(\beta)}}{\sum_{k\in C_n} e^{V_{nk}(\beta)}}\right)$$

We begin by noting that we can assume that the characteristics of each individual, and the choice set of alternatives associated with each individual are all drawn from a data generating process defined by a random distribution. Let

$$LM(\beta,m) = \ln\left(\frac{e^{V_{mj}(\beta)}}{\sum_{k\in C_n} e^{V_{mk}(\beta)}}\right)$$

where j is a random variable representing the alternative chosen by household n. If these draws are independent, then LM_N represents a sum of independent and identically distributed random variables $\sum_{n=1}^{N} LM(\beta,n)$ and therefore, we can apply the law of large numbers concluding that the limit of the probability distribution of LM_N is in fact $E(LM(\beta,m))$.

The term $LM(\beta, m)$ depends on three random variables, the household m, the choice set C drawn for household m, and the choice of alternative, j. This means that

$$E(LM(\beta,m)) = \int_{j,C,m} \ln\left(\frac{e^{V_{mj}(\beta)+\ln P(C|j,m)}}{\sum_{k\in C} e^{V_{mk}(\beta)+\ln P(C|k,m)}}\right) P(j,C,m)djdDdm$$

$$= \int_{j,C,m} \ln\left(\frac{P(C|j,m)e^{V_{mj}(\beta)}}{\sum_{k\in C} P(C|k,m)e^{V_{mk}(\beta)}}\right) P(C|j,m)P(j|m)P(m)djdDdm \tag{2.2}$$

$P(C|j,m)$ and $P(j|m)$ are both discrete distributions, and assuming that the choice of alternative made by household m conforms to the Conditional Logit model, then

$$P(j|m) = \frac{e^{V_{mj}(\beta^*)}}{\sum_{k\in U} e^{V_{mk}(\beta^*)}} \tag{2.3}$$

where β^* are the true model parameters. The estimators of the maximum likelihood parameters, and the estimators of the modified maximum likelihood parameters will be consistent, if the modified maximum likelihood estimates converge to β^* as N goes to infinity. To save space, we replace $P(C|k,m)$ with π_k. Replacing this equation in 2.2

and observing that π_k is zero when $k \notin C$ we obtain

$$
\begin{aligned}
& E(LM(\beta, m)) \\
&= \int_m \sum_{j \in U} \sum_{C \subset U} \ln \left(\frac{\pi_j e^{V_{mj}(\beta)}}{\sum_{k \in U} \pi_k e^{V_{mk}(\beta)}} \right) \frac{e^{V_{mj}(\beta^*)}}{\sum_{k \in U} e^{V_{mk}(\beta^*)}} \pi_j P(m) dm \\
&= \int_m \sum_{C \subset U} \sum_{j \in U} \frac{e^{V_{mj}(\beta^*)}}{\sum_{k \in U} e^{V_{mk}(\beta^*)}} \frac{\sum_{k \in U} \pi_k e^{V_{mk}(\beta^*)}}{\sum_{k \in U} \pi_k e^{V_{mk}(\beta^*)}} \ln \left(\frac{\pi_j e^{V_{mj}(\beta)}}{\sum_{k \in U} \pi_k e^{V_{mk}(\beta)}} \right) \pi_j P(m) dm \\
&= \int_m \sum_{C \subset U} \frac{\sum_{k \in U} \pi_k e^{V_{mk}(\beta^*)}}{\sum_{k \in U} e^{V_{mk}(\beta^*)}} \sum_{j \in U} \frac{\pi_j e^{V_{mj}(\beta^*)}}{\sum_{k \in U} \pi_k e^{V_{mk}(\beta^*)}} \ln \left(\frac{\pi_j e^{V_{mj}(\beta)}}{\sum_{k \in U} \pi_k e^{V_{mk}(\beta)}} \right) P(m) dm
\end{aligned}
$$

Letting $\phi(\beta, j)$ equal $\frac{\pi_j e^{V_{mj}(\beta)}}{\sum_{k \in U} \pi_k e^{V_{mk}(\beta)}}$, we have

$$
E(LM(\beta, m)) = \int_m \sum_{C \subset U} \frac{\sum_{l \in U} \pi_l e^{V_{ml}(\beta^*)}}{\sum_{k \in U} e^{V_{mk}(\beta^*)}} \sum_{j \in U} \phi(\beta^*, j) \ln \phi(\beta, j) P(m) dm
$$

In order to show that $E(LM(\beta, m))$ achieves its maximum where β equals β^*, we need only show that $\sum_{j \in U} \phi(\beta^*, j) \ln \phi(\beta, j)$ is maximized at this value of β. We start by finding the stationary point of this expression.

$$
\frac{\partial}{\partial \beta} \sum_{j \in U} \phi(\beta^*, j) \ln \phi(\beta, j) = \sum_{j \in U} \frac{\phi(\beta^*, j)}{\phi(\beta, j)} \frac{\partial}{\partial \beta} \phi(\beta, j)
$$

If $\beta = \beta^*$, then noting that $\sum_{j \in U} \phi(\beta, j) = 1$,

$$\begin{aligned}
\left[\frac{\partial}{\partial \beta} \sum_{j \in U} \phi(\beta^*, j) \ln \phi(\beta, j) \right]_{\beta = \beta^*} &= \sum_{j \in U} \frac{\partial}{\partial \beta^*} \phi(\beta^*, j) \\
&= \frac{\partial}{\partial \beta^*} \left(\sum_{j \in U} \phi(\beta^*, j) \right) \\
&= \frac{\partial}{\partial \beta^*} 1 \\
&= 0
\end{aligned}$$

And thus β^* is a stationary point. We now need to prove that this stationary point is a global maximum. We do this by proving that this expression has a negative definite Hessian matrix. This proof follows the same form as that found in Appendix 2

$$\begin{aligned}
&\frac{\partial^2}{\partial \beta^2} \sum_{j \in U} \phi(\beta^*, j) \ln \phi(\beta, j) \\
&= \sum_{j \in U} \phi(\beta^*, j) \frac{\partial}{\partial \beta} \left(X_{nj} - \sum_{k \in U} \phi(\beta, k) X_{nk} \right)^T \\
&= \sum_{j \in U} -\phi(\beta^*, j) \sum_{k \in U} \frac{\partial}{\partial \beta} \phi(\beta, k) X_{nk}^T \\
&= \sum_{j \in U} -\phi(\beta^*, j) \sum_{k \in U} \left(X_{nk} - \sum_{l \in U} \phi(\beta, l) X_{nl} \right) \phi(\beta, k) X_{nk}^T \\
&= \sum_{j \in U} \sum_{k \in U} -\phi(\beta^*, j) \phi(\beta, k) X_{nk} X_{nk}^T \\
&+ \sum_{j \in U} \sum_{k \in U} \sum_{l \in U} \phi(\beta^*, j) \phi(\beta, k) \phi(\beta, l) X_{nl} X_{nk}^T \\
&= \sum_{j \in U} \sum_{k < l} -\phi(\beta^*, j) \left(\phi(\beta, k) X_{nk} - \phi(\beta, l) X_{nl} \right) \left(\phi(\beta, k) X_{nk} - \phi(\beta, l) X_{nl} \right)^T
\end{aligned}$$

Which is negative definite. The last step is derived in much the same way as in the derivation of the proof in Equation 4.2 of Appendix 2. We have thus shown that the

value of β that maximizes the probability limit of $LM_N(\beta)$ is equal to the parameters of the probability limit of the true model likelihood L_N. However, we have not shown that the values of β that maximize $LM_N(\beta)$ converge in probability to the maximum of the probability limit of $LM_N(\beta)$. This is shown in a proof that appears in Manski and McFadden (1977 [30]).

With these results, we now have a way of reliably approximating the maximum likelihood estimates of a fully specified Logit model while vastly reducing the cost of calculation. We now investigate a few choice set assignment mechanisms, checking for the Positive and Uniform Conditional Properties in each case.

Examples of choice set assignment mechanisms

We consider here several different assignment mechanisms, evaluating whether they satisfy the positive conditioning and uniform conditioning properties.

Universal choice set

This is the degenerate case, in which we take as the choice set for every individual, the entire choice set, U, so that the probability distribution of assignment of choice sets C is the following:

$$\pi(C|n,j) = \begin{cases} 1, \text{ if } C = U \\ 0, \text{ otherwise} \end{cases}$$

giving us a probability

$$\begin{aligned} P_{nj}(\beta) &= \frac{e^{V_{nj}(\beta)+\log \pi(C|n,j)}}{\sum_{k\in C} e^{V_{nk}(\beta)+\log \pi(C|n,k)}} \\ &= \frac{e^{V_{nj}(\beta)}}{\sum_{k\in U} e^{V_{nk}(\beta)}} \end{aligned}$$

Note that this assignment mechanism is in fact independent of n and j. The verification of the positive conditioning property, the uniform conditioning property, and the consistency of the model parameters is trivial.

Simple Random Sampling of alternatives

In this section, we attempt to create an assignment mechanism that adheres to both the positive and uniform conditioning properties. The first technique is the simplest. We include the observed alternative in C, and then we select a subset $C^S_{n,j}$ of $U - \{j\}$ using J draws, sampling without replacement, and with equal probabilities of selection for all stores.

The conditional probability of drawing such a set will be

$$\begin{aligned} \pi(C|n,j) &= \pi(C^S_{n,j}|n,j) \\ &= \begin{cases} \left[\binom{|U|-1}{J}\right]^{-1} & \text{if } j \in C \text{ and } |C| = J+1 \\ 0, \text{ otherwise} \end{cases} \end{aligned} \tag{2.4}$$

Thus, provided that $i, j \in C$ and $|C| = J + 1$,

$$\pi(C|n,i) = \pi(C|n,j) = \frac{J!(|U|-1-J)!}{(|U|-1)!} \tag{2.5}$$

and the uniform conditioning property, and thus the positive conditioning properties are both verified.

Sampling proportional to relevancy of alternatives

Because we know that households are far more likely to select stores that are close by than far away, we would like our assignment mechanism to reflect this, in order to increase

the expected number of likely alternatives included in our choice set. In this technique, every alternative is assigned a strictly positive sampling weight, W, proportional to the relevance of the given alternative. This will probably be related to the gravitational attraction of the store (i.e., related to an inverse function of the distance of the store) or it could be something else. We draw J alternatives from the set of alternatives U and add to this selection the alternative chosen by the given household, but this time, we follow the sample weights. This sample of stores can be chosen with or without replacement. We consider both cases.

Sampling with replacement

In this method, also described by Ben-Akiva and Lerman (1985 [6])we construct the choice set C_n for a given individual n by making J independent draws with replacement from U, using the weighting defined above, adding the observed store choice for individual n, and then eliminating all duplicate alternatives. Using this method, we do not know the number of alternatives in C_n in advance, except that it must be between 1 and $J+1$.

Let

$$q_j = \frac{W_j}{\sum_{k \in U} W_k}$$

Where the weight W_j represents the importance of the alternative j. The probability of constructing C_n, if C_n contains $J+1$ elements is

$$P(C_n|j, J+1 \text{ elements}) \quad = \quad J! \prod_{k \in C_n, k \neq j} q_k$$

since none of the draws can be duplicates or equal to j. The probability of constructing C_n if there are J elements is the same as the probability of making $J-1$ draws with no duplicates, and drawing an alternative that is a duplicate at any of the J possible points

(first, second, etc) within the draw. This gives us

$$P(C|n, j, J \text{ elements}) = (J-1)! \prod_{k \in C_n, k \neq j} q_k \left(J \sum_{l \in C_n} q_l \right)$$

By induction, we can show that

$$P(C|n, j) = J! \prod_{k \in C_n, k \neq j} q_k \left(\sum_{l \in C_n} q_l \right)^{J-|C_n|+1}$$

where $|C_n|$ is the number of elements in C_n. This is equal to

$$\begin{aligned} P(C|n, j) &= q_j^{-1} J! \prod_{k \in C_n} q_k \left(\sum_{l \in C_n} q_l \right)^{J-|C_n|+1} \\ &= q_j^{-1} Q_n \end{aligned}$$

where Q_n is an expression that does not depend on j. We see that the positive conditioning property is verified, but not the uniform conditioning property, since $P(C|n, j) = P(C|n, i)$ is equivalent to $q_j = q_i$, since Q_n cancels out on both sides of the equation. Thus, the uniform conditional property is satisfied if and only if $q_j = q_i$ for all i and j in C_n. The positive conditioning property is satisfied if and only if the alternative chosen by n is assigned a positive weight (and therefore q_j is nonzero).

When we replace this into our modified expression for the estimated probabilities of selection, Q_n appears on both the numerator and denominator of the expression and cancels out, so that we have:

$$\begin{aligned} P(j|C, n) &= \frac{P(C|j, n) e^{V_{nj}}}{\sum_{k \in C} P(C|k, n) e^{V_{nk}}} \\ &= \frac{q_j^{-1} e^{V_{nj}}}{\sum_{k \in C} q_k^{-1} e^{V_{nk}}} \end{aligned}$$

This is a simple expression to calculate, and is the form suggested by Ben-Akiva and Lerman (1985 [6]). The only drawback with this approach is that the number of alternatives in each choice set is not known before a draw is undertaken.

Sampling without replacement

Let C_n be determined for household n by including the chosen alternative and adding J alternatives drawn from the set $U - \{j\}$ without replacement. Let $C^S_{n,j}$ be the set $C_n - \{j\}$. If j is the chosen alternative, then $P(C|n,j)$ will be the probability of selecting the set $C^S_{n,j} = \{k_1, \ldots, k_J\}$ from $U - \{j\}$ without replacement. As in sampling with replacement, let

$$q_j = \frac{W_j}{\sum_{k \in U} W_k}$$

The probability of selecting the set of alternatives $C^S_{n,j}$ without replacement from $U - \{j\}$ in the order $k_1, k_2, \ldots, k_J$ will be defined as

$$\begin{aligned} P(C^S_{n,j}|O_{1,j}) &= \frac{q_{k_1}}{1 - q_j} \frac{q_{k_2}}{1 - q_j - q_{k_1}} \frac{q_{k_3}}{1 - q_j - q_{k_1} - q_{k_2}} \cdots \frac{q_J}{1 - q_j - \sum_l^{J-1} q_{k_l}} \\ &= \prod_s^J \frac{q_{k_s}}{1 - q_j - \sum_{l=1}^{s-1} q_{k_l}} \end{aligned}$$

where $O_{1,j}$ is the ordered set of elements of $C^S_{n,j}$ arranged according to the sequence

in which these alternatives are drawn. The possible ordered sets for $C_{n,j}^S$ are as follows:

$$\begin{aligned}
O_{1,j} &= \{k_1, k_2, k_3, \ldots, k_{J-2}, k_{J-1}, k_J\} \\
O_{2,j} &= \{k_1, k_2, k_3, \ldots, k_{J-2}, k_J, k_{J-1}\} \\
O_{3,j} &= \{k_1, k_2, k_3, \ldots, , k_J, k_{J-2}, k_{J-1}\} \\
O_{4,j} &= \{k_1, k_2, k_3, \ldots, k_J, k_{J-1}, k_{J-2}\} \\
\vdots &\quad \vdots \\
O_{J!-1,j} &= \{k_J, k_{J-1}, k_{J-2}, \ldots, k_3, k_1, k_2\} \\
O_{J!,j} &= \{k_J, k_{J-1}, k_{J-2}, \ldots, k_3, k_2, k_1\}
\end{aligned}$$

Let $k_t(s,j)$ be the t^{th} term of $O_{s,j}$. Thus, $k_t(1,j) = k_t$, $k_J(2,j) = k_{J-1}$, and $k_1(J!,j) = k_J$. Let $P(k_t(s,j)|O_{s,j})$ be the probability of selecting $k_t(s,j)$ in the t^{th} position when the elements of $C_{n,j}^S$ are drawn in the order defined by $O_{s,j}$. Then

$$P(k_t(s,j)|O_{s,j}) = \frac{q_{k_t(s,j)}}{1 - q_j - \sum_{v=1}^{t-1} q_{k_v(s,j)}}$$

If we define

$$S_t(s,j) = \sum_{v=1}^{t} q_{k_v(s,j)}$$

then

$$P(k_t(s,j)|O_{s,j}) = \frac{q_{k_t(s,j)}}{1 - q_j - S_{t-1}(s,j)}$$

We can therefore write

$$P(C|n,j) = \sum_{s=1}^{J!} P(C^S_{n,j}|O_{s,j})$$

and taking $S_0(s,j)$ to be zero,

$$\begin{aligned} P(C^S_{n,j}|O_{s,j}) &= P(k_1(s,j)|O_{s,j})P(k_2(s,j)|O_{s,j})P(k_3(s,j)|O_{s,j})\cdots P(k_J(s,j)|O_{s,j}) \\ &= \frac{q_{k_1(s,j)}}{1-q_j}\frac{q_{k_2(s,j)}}{1-q_j-S_1(s,j)}\frac{q_{k_3}(s,j)}{1-q_j-S_2(s,j)} \\ &\quad \cdots \frac{q_{k_J}(s,j)}{1-q_j-S_{J-1}(s,j)} \\ &= \prod_{t=1}^{J}\frac{q_{k_t(s,j)}}{1-q_j-S_{t-1}(s,j)} \end{aligned}$$

The Positive Conditioning Property of this strategy is established whenever W_k is nonzero for k in C_n. We can show by counterexample that the Uniform Conditioning Property does not generalize over sampling without replacement when there are unequal sampling weights. Supposing that we have a choice set C_n equal to $\{a,b,c\}$. Suppose that we calculate that $q_a = 1/4$, $q_b = 1/5$, and $q_c = 1/10$. If a were the drawn alternative, then $C^S_{n,a} = \{b,c\}$. Then

$$\begin{aligned} P(C|n,a) &= \left(\frac{1/5}{1-1/4}\right)\left(\frac{1/10}{1-1/4-1/5}\right) + \left(\frac{1/10}{1-1/4}\right)\left(\frac{1/5}{1-1/4-1/10}\right) \\ &= 0.08951 \end{aligned}$$

Likewise,

$$\begin{aligned} P(C|n,b) &= \left(\frac{1/4}{1-1/5}\right)\left(\frac{1/10}{1-1/5-1/4}\right) + \left(\frac{1/10}{1-1/5}\right)\left(\frac{1/4}{1-1/5-1/10}\right) \\ &= 0.10146 \end{aligned}$$

We see here that although sampling without replacement has the advantage of allowing us to determine choice set sizes in advance of the alternative sampling, the calculation of the probabilities of drawing samples is excessively complicated with respect to sampling with replacement, considering we have to calculated a new conditional probability for every single order of alternative draws, something that could cause calculations to become quickly intractable.

Mixed deterministic and random choice set assignment

We also consider devising a strategy in which we use a nonrandom procedure to assign some alternatives to a choice set, and then do J draws from the remaining choices in the universe of choice to obtain an additional set of alternatives. Once we obtain these two disjoint sets, we combine them and add the store chosen by the household to obtain the choice set of the household. Thus, if $\{j\} \in C$ then

$$C = C'_n \cup C^S_n \cup \{j\}$$

where C'_n is the set of nonrandomly assigned stores to household n, and C^S_n is a sample of stores selected at random from $U - C'_n$ (so that $C'_n \cap C^S_n = \emptyset$). Adding nonrandom selections to our choice set does not fundamentally alter the way we calculate the draw probabilities of our choice sets, since we need only replace U with $U - C'_n$ in the derivations in the previous sections. If we created a choice set C containing the alternatives i and j, and we were to draw the alternatives at random with replacement, then we would

obtain, with $q_j = \frac{W_j}{\sum_{k \in C_n - C'_n} W_k}$ and $Q_n = J! \prod_{k \in C_n^S} q_k \left(\sum_{l \in C_n^S} q_l \right)^{J - |C_n^S| + 1}$:

$$P(C|n,j) = \begin{cases} q_j^{-1} Q_n \text{ if } j \notin C'_n \\ Q_n \text{ if } j \in C'_n \end{cases}$$

if j were the chosen store, and

$$P(C|n,i) = \begin{cases} q_i^{-1} Q_n \text{ if } i \notin C'_n \\ Q_n \text{ if } i \in C'_n \end{cases}$$

if i were. Once again, we see that the positive conditioning property is satisfied, however, since some alternatives in C are drawn, and some are not, then the uniform conditioning property cannot be satisfied, even when we assign equal sampling weights to all sampled alternatives.

In the case where we use sampling without replacement, we use the following redefinition:

$$P(C_{n,j}^S | O_{s,j}) = W_{s \in U - C'_n} \frac{q_{k_s}}{1 - q_j - \sum_{l=1}^{s-1} q_{k_l}}$$

with $O_{s,j}$ redefined to include no more individuals than are in $|C'_n|$.

$$P(C|n,j) = \begin{cases} \sum_{s=1}^{J!} P(C_{n,j}^S | O_{s,j}) \text{ if } j \notin C'_n \\ \sum_{s=1}^{J!} P(C_n^S) \text{ if } j \in C'_n \end{cases}$$

if j were the chosen store, and

$$P(C|n,i) = \begin{cases} \sum_{s=1}^{J!} P(C_{n,i}^S|O_{s,i}) \text{ if } i \notin C_n' \\ \sum_{s=1}^{J!} P(C_n^S) \text{ if } i \in C_n' \end{cases}$$

if i were. We have already shown that if sampling weights are unequal, then if $i, j \in C_{n,j}^S$ then $P(C_{n,i}^S|O_{s,i}) \neq P(C_{n,j}^S|O_{s,j})$. If sampling weights were equal, then we would be in the case where sampled alternatives were drawn through simple random sampling, and thus, if $C_{n,i}^S$ contained J alternatives,

$$P(C|n,j) = \begin{cases} \frac{(J-1)!(|U|-|C'|-1-J)!}{(|U|-|C'|-1)!} \text{ if } j \notin C_n' \\ \frac{(J)!(|U|-|C'|-J)!}{(|U|-|C'|)!} \text{ if } j \in C_n' \end{cases}$$

if j were the chosen store, and

$$P(C|n,i) = \begin{cases} \frac{(J-1)!(|U|-|C'|-1-J)!}{(|U|-|C'|-1)!} \text{ if } i \notin C_n' \\ \frac{(J)!(|U|-|C'|-J)!}{(|U|-|C'|)!} \text{ if } i \in C_n' \end{cases}$$

if i were. Thus, for a combination of a fixed choice set and sampling without replacement, the Positive Conditioning, but not the Uniform Conditioning Properties are satisfied.

Remarks

The use of sampled alternatives is an elegant solution to the problem of large choice sets, but must be regarded with caution. The consideration sets here have no resemblance to the real set of alternatives considered by the individuals in question. Moreover, the introduction of random draws of alternatives could introduce an additional source of random variation that will reduce the reliability of a model's adjustment to a finite sample of households, thus negating any advantages this model might have in terms of

asymptotic qualities.

2.2 Evaluating the predictive accuracy

In this section, we shall look at how we shall evaluate our conditional logit model. There are two main ways in which to evaluate a qualitative response model, either through model fit statistics that measure the fit of a model's predicted probabilities and observed response frequencies, or through measures of the ability of a model to forecast observed responses. In the first two parts of the section, we shall look at a classic measure of model fit, the McFadden Pseudo R^2 index, and a classic measure of response accuracy, the Brier Score. In the third part, we shall describe the measure that we develop in order to evaluate our models. Our measure of model fit is based on the intended use of the probabilities of store selection produced by our model to generate accurate predictions of market shares of stores within predefined geographic zones in France. We have developed an index that provides an intuitive way of evaluating the ability of our model to produce results that work well in this specific application.

2.2.1 The McFadden Pseudo R^2

Since the log likelihood was the criteria by which we determined the best model parameter estimates, it would be logical to use the log likelihoods in order to judge the overall quality of the model. The McFadden Pseudo R^2 (proposed in McFadden, 1974a [32]) is an index that measures the increase in the log likelihood of a model with respect to a null log likelihood.

If we define the maximum log likelihood estimate of the fitted model over a sample

S of size N as in 1.1

$$\begin{aligned} \mathrm{LL}(\beta) &= \mathrm{argmax}_\beta \left(\sum_{n=1}^{N} \sum_{j \in C_n} z_{nj} \ln P_{nj} \right) \\ &= \mathrm{argmax}_\beta \left(\sum_{n=1}^{N} \sum_{j \in C_n} z_{nj} \ln \left(\frac{e^{V_{nj}(\beta)}}{\sum_{k \in C_n} e^{V_{nk}(\beta)}} \right) \right) \end{aligned}$$

and the log likelihood of a null model in which utilities are determined only by alternative-specific constants as

$$\begin{aligned} \mathrm{LL}_0(\beta) &= \mathrm{argmax}_\beta \left(\sum_{n=1}^{N} \sum_{j \in C_n} z_{nj} \ln P^0_{nj} \right) \\ &= \mathrm{argmax}_\alpha \left(\sum_{n=1}^{N} \sum_{k \in C_n} z_{nj} \ln \left(\frac{e^{\alpha_j}}{\sum_{k \in C_n} e^{\alpha_k}} \right) \right) \end{aligned}$$

with α representing choice-specific constants, then the McFadden R^2 term is defined as:

$$\text{McFadden } \mathrm{R}^2 = 1 - \frac{\mathrm{LL}}{\mathrm{LL}_0}$$

We note that LL_0 term can be simplified to

$$\mathrm{LL}_0(\beta) = \sum_{n=1}^{N} \sum_{k \in C_n} \ln \frac{n_k}{N_k}$$

where n_k is the number of individuals in N who choose k and N_k is the number of individuals in S for whom k is an alternative within their choice set. We observe that

$$\begin{aligned} P^0_{nj} &= \frac{e^{\alpha_j}}{\sum_{k \in C_n} e^{\alpha_k}} \\ &= \frac{e^{\alpha_1 x_{nj1} + \cdots + \alpha_J x_{njJ}}}{\sum_{k=1}^{J} e^{\alpha_1 x_{nk1} + \cdots + \alpha_J x_{nkJ}}} \end{aligned}$$

where x_{nkl} is one when $k = l$ and $k \in C_n$ and zero elsewhere. We suppose without loss of generality that $U = \bigcup_n C_n = \{1, \ldots, J\}$. From Equation 4.3 in Appendix 4, we find that the maximum likelihood estimates of α, the vector of parameters $(\alpha_1, \ldots, \alpha_J)^T$ is, with X_{nk} the vector of indicator variables $(x_{nk1}, \ldots, x_{nkJ})^T$

$$\sum_{n=1}^{N} \sum_{k=1}^{J} (z_{nk} - P^0_{nk}) X_{nk} \quad = \quad 0 \tag{2.6}$$

which implies, (since $\sum_{k=1}^{J} a_k x_{nk} = a_k$), that

$$\sum_{n \in S_k} z_{nk} \quad = \quad \sum_{n \in N_k} P^0_{nk}$$

for all choices k, where S_k is the set of households for which k is an alternative in the choice set of n. Since P^0_{nk} is constant,

$$P^0_{nk} \quad = \quad \frac{n_k}{N_k}$$

.

The McFadden R^2 is designed to resemble the familiar R^2 of ordinary linear regression, which is defined

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

where SSE represents the sum of squared residual terms, and SST represents the total variation of the dependent variable. This term corresponds to the proportion of overall dependent variable variation that is accounted for by the model's predictions. The

SST term is the upper bound on SSE that is attained when no auxiliary information is introduced into the model and the average dependent variable responses are assigned as the predicted responses of each individual. If we prefer to use the log likelihood of our model to evaluate our model instead of the sum of squared residuals, $-\mathrm{LL}$ and $-\mathrm{LL}_0$ have properties analogous to SSE and SST. $-\mathrm{LL}_0$ is the upper bound of $-\mathrm{LL}$ that is attained when no auxiliary information is entered into the conditional logit model and every individual is assigned the sample average probability of selection of each alternative in their choice set as their own probability of selection. Unlike in ordinary R^2, unfortunately, the value of the McFadden R^2 does not represent the actual percentage by which the variation in the dependent variable is reduced by using the model, and some simulation studies have shown that a value as low as 0.30 corresponds to high values of ordinary R^2 terms (pp 134-135, Domencich and McFadden, 1975 [11]).

One advantage of this index, as Domencich and McFadden note, is that it can serve as a test of the significance of model parameters. If K is the number of parameters in the model, and K' is the number of constants within the model parameters, R^2 is a random variable such that on large samples,

$$\frac{K}{K-K'}\frac{R^2}{1-R^2} \quad \sim \quad F(K-K', K)$$

This can be used to test the probability that the parameters used to calculate LL_1 are the same as those in LL_0, that is, that the nonconstant explanatory variables are significant. This is actually the familiar likelihood ratio test. We can generalize this technique to use any set of parameters to define the log likelihood value for LL_1, not necessarily the full set of model parameters, and any set of parameters to define the log likelihood value for the null model, provided that the null model parameters are contained within the fitted model parameters. By using this test, we can see if the

addition of parameters to any model will significantly improve the log likelihood.

What interests us in our paper is not so much the fit of a model's predicted probabilities to the data set on which it is based, but the accuracy of the model's predictions when applied to a new data set. Although interesting, the McFadden R^2 test will not be our basis for model evaluation.

2.2.2 Brier Score

In this section, we propose the Brier Score (proposed in Brier, 1950 [7]) as an indicator of predictive accuracy, but we show that this indicator has limited usefulness when our model does not take into account individual variations. This score, which is mainly used in the field of meteorology, applies to predictions of a binary response variable. If z refers to this variable, then the Brier Score is defined as the average squared difference between the observed values of the variable on a sample S of N individuals, and the predicted probability of obtaining positive responses P.

$$\text{Brier Score} \quad = \quad \frac{1}{N}\sum_{n=1}^{N}(z_n - P_n)^2$$

The Brier Score is between zero and one, one being the worst-case scenario where a zero probability is attributed to every positive response and a probability of one to every negative response, while a Brier Score of zero corresponds to the best-case scenario where the opposite is the case. A Brier Score can be applied to the multinomial case by taking every single alternative for every single individual in our choice set as a separate observation and defining a binary variable z_{nk} that is one when n chooses k and zero,

otherwise. This gives the following formula.

$$\text{Brier Score} \;=\; \frac{1}{2N}\sum_{n=1}^{N}\sum_{k\in C_n}(z_{nk}-P_{nk})^2$$

We do not divide by the number of observations as we do in the binary case, since the values of P_{nk} and z_{nk} are constrained to sum up to one for every n. Instead, we divide by $2N$, since in a worst-case scenario, for every n, P_{nk} would be one for one of the nonchosen alternatives, and zero for all other alternatives, making $\sum_{n=1}^{N}\sum_{k\in C_n}(z_{nk}-P_{nk})^2$ equal to $2N$. In a best-case scenario, the Brier Score would still be zero, as in the binary case. By taking n to be households and k to be alternatives in each households' choice of stores, this modified Brier Score obviously can be applied to our data.

We need to express a word of caution, however. Although the Brier Score may give an idea of how close our predictions are to the observations, it may not be a completely meaningful judge of a model's quality. First of all, this score depends on the proportion of positive observations in our population. If we take a binary model as an example, a Brier Score of 0.10 must be interpreted much differently for a sample that has a rate of positive response of 0.50 percent than a sample that has a rate of positive response of 0.95. Although in the first case, a Brier Score of 0.10 may seem good, this is not the case for the second. In the first case, if we simply assigned a probability of 1 to every single response in the sample, we would have a Brier Score of 0.50, but in the second, the Brier Score would be 0.05, which is a better score than the one produced by the proposed model.

We find that Brier Score is a quantity that depends on the distribution of observed responses, the distribution of the model predictions, and the interaction between the observed responses and model predictions. The more variation there is in the variable we wish to predict, the more it will differ from a null model in which the predicted

response is represented by a constant. In order to obtain an index of model predictive quality, we will need to be able to calculate how much of the Brier Score is simply a reflection of the inherent variability of the response variable, or, the "Uncontrolled Variation" of the variable. The Uncontrolled Variation should work as an upper limit to the Brier Score that remains independent of any predictions that we may make.

We can consider the assigning of predictions to individuals to consist of two steps. In the first, we identify the factors that are related to the values of our response variable and attempt to represent these with explanatory variables. Then, we quantify the factors we have identified by assigning coefficients to our explanatory variables in order to assign predictions to each individual. We would like to be able to evaluate how successfully we accomplish each of these tasks separately.

When the values of the explanatory variables do not differ between two individuals, the predicted values assigned to each individual do not differ either and therefore the difference that remains in the response variable represents the variability of the response variable that cannot be taken into account by our model, no matter how accurately we may calculate the coefficients of our explanatory variables. If we divide our population into groups with respect to the values of the explanatory variables we use, then the between-group variation of our response variable will represent the maximum amount of uncontrolled variation that can be accounted for by our model. This quantity, we shall call the "Resolution", and it is a reflection of our ability to identify model effects. The greater the Resolution in our model, the better the Brier Score can potentially be.

The "Calibration" of our model refers to the accuracy of the predictions that we assign to our individuals. This value does not take into account the variation between responses assigned the same predicted value, but instead calculates the squared difference between each prediction, and the average value of the response value for all individuals assigned

the same prediction. The greater the calibration, the higher the Brier Score, and the worst our model estimations are. Whereas the Resolution is a measure of the quality of a model specification, the Calibration is a measure of the quality of the model's parameter estimates.

Suppose that S is partitioned into G distinct groups of individual-alternative pairs (n,k) for which P_{nk} is constant. That is, if g is in G, $P_{nk} = P_g$ for all (n,k) in g. We define z_g as the average of z_{nk} in g, $\bar{z}$ as the average of z_{nk} over the entire population, N as the number of households in the entire sample. This allows us to obtain the formal definitions of the three indexes we defined above:

$$\text{Uncontrolled Variation} = \frac{1}{2N}\sum_{g=1}^{G}\sum_{(n,k)\in g}(z_{nk}-\bar{z})^2$$

$$\text{Resolution} = \frac{1}{2N}\sum_{g=1}^{G}\sum_{(n,k)\in g}(z_{g}-\bar{z})^2$$

$$\text{Calibration} = \frac{1}{2N}\sum_{g=1}^{G}\sum_{(n,k)\in g}(z_{g}-P_g)^2$$

It is easily shown that these three indexes are in fact different components of the Brier Score.

Theorem 2 (The Murphy Decomposition (1972))

$$\textit{Brier Score} = \textit{Uncontrolled Variation} - \textit{Resolution} + \textit{Calibration}$$

Proof:

$$
\begin{aligned}
\frac{1}{2N} \quad & \sum_{n=1}^{N} \sum_{k \in C_n} (z_{nk} - P_{nk})^2 \\
& = \frac{1}{2N} \sum_{g=1}^{G} \sum_{(n,k) \in g} (z_{nk} - P_g)^2 \\
& = \frac{1}{2N} \sum_{g=1}^{G} \sum_{(n,k) \in g} (z_{nk} - z_g + z_g - P_g)^2 \\
& = \frac{1}{2N} \sum_{g=1}^{G} \sum_{(n,k) \in g} \left[(z_{nk} - z_g)^2 + (z_g - P_g)^2 + (z_{nk} - z_g)(z_g - P_g) \right] \\
& = \frac{1}{2N} \sum_{g=1}^{G} \sum_{(n,k) \in g} (z_{nk} - z_g)^2 + \frac{1}{2N} \sum_{g=1}^{G} \sum_{(n,k) \in g} (z_g - P_g)^2 \qquad (2.7)
\end{aligned}
$$

Since P_g, z_g, and $\bar{z}$ are constant over the groups g, the sum of the cross terms is zero. Similarly,

$$
\sum_{g=1}^{G} \sum_{(n,k) \in g} (z_{nk} - \bar{z})^2 \;=\; \sum_{g=1}^{G} \sum_{(n,k) \in g} (z_{nk} - z_g)^2 + \sum_{g=1}^{G} \sum_{(n,k) \in g} (z_g - \bar{z})^2
$$

Substituting this in 2.7, we obtain that the Brier Score can be written as:

$$
\frac{1}{2N} \sum_{g=1}^{G} \sum_{(n,k) \in g} \left[(z_{nk} - \bar{z})^2 - (z_g - \bar{z})^2 + (z_g - P_g)^2 \right]
$$

■

The Uncontrolled Variation depends only on the proportion of positive observations and is independent of any modelling that we may use. The resolution depends only on the definition of the groups g in the model and is independent of the probabilities assigned by our model. This means that when we compare the predictions made by

Zone	Brier	UV	Resol	Calib
Ind	0.355	0.439	0.439	0.335
IRIS	0.355	0.439	0.175	0.071

Table 2.1: Murphy Decomposition by individual or by "IRIS"

different models, if there is a set of groups G for which we know that our model predicted probabilities remains constant, a comparison of the calibrations of the different models will be just equivalent to a comparison of the Brier Scores. Because in our predictions of store choice, we have not taken into account sociodemographic characteristics that distinguish between individual households, the predicted probabilities will be constant for domiciles having the same geographic co-ordinates, which, in our case, are households living in the same census district, or "IRIS". This means that we can define our Murphy Decomposition of the Brier Score with each g as containing a unique household-store choice pair on the one hand, or we can define each group g such that if $(n, k) \in g$, then (m, k) will be in g if m and n are in the same "IRIS". Doing this, we obtain the results in Table 2.1:

We can see that in the first line, where each g corresponds to a unique observation, the Calibration is simply the Brier Score, and the Uncontrolled Variation and Resolution cancel out. However, in the second line, we see that although the Brier Score and the Uncontrolled Variation remain the same, the resolution and calibration have both changed. The second calibration is perhaps more meaningful than the first, for it takes into account the level of detail of our model's predictions. We can also see that the calibration in the first case is inflated by within-"IRIS" variation in observed responses that causes a much higher resolution term, and therefore a higher calibration term to compensate. There is no point in taking this into account in comparing different models that do not take this into account either.

We note that if Calibration is

$$\frac{1}{2N}\sum_{g=1}^{G}\sum_{(n,k)\in g}(z_g - P_g)^2$$

we are looking at the distances between aggregated predictions of households on population subgroups g. It has been observed at BVA that perhaps a higher level of aggregation would be more meaningful, since the sample of households is not designed to be representative at the "IRIS" level, but at the level of survey sectors. Therefore, instead of comparing estimates of the probability of a given household from each "IRIS" visiting each store, we ought to be looking at estimates of the average probability of a household from a given sector visiting a given store. In that case, g will represent a set of household-store-choice pairs in which the household associated with every pair is in the same geographic zone, and each pair includes the same store choice. With this new definition of g, z_g represents the average response as before, but P_g represents the average predicted probability in g. If we prefer the new calibration, replacing P_{nk} with P_g no longer leaves the Brier Score unchanged. In fact, we find that if we define BS as the Brier Score of a set of predicted probabilities, and BS(G) as the Brier Score of the averages of the predicted probabilities for the groups in G, so that

$$\text{BS} \;=\; \frac{1}{2N}\sum_{g=1}^{G}\sum_{(n,k)\in g}(z_{nk} - P_{nk})^2$$

and

$$\text{BS}(G) \;=\; \frac{1}{2N}\sum_{g=1}^{G}\sum_{(n,k)\in g}(z_{nk} - P_g)^2$$

we obtain the following simple result:

Theorem 3 *If var$(P|G)$ represents the within-group variation of P, and cov$(z, P|G)$ represents the average covariance of z and P for each group g in G, then*

$$BS = BS(G) + var(P|G) - 2cov(z, P|G)$$

Proof:

$$\begin{aligned} \text{BS} &= \frac{1}{2N}\sum_{g=1}^{G}\sum_{(n,k)\in g}(z_{nk} - P_{nk})^2 \\ &= \frac{1}{2N}\sum_{g=1}^{G}\sum_{(n,k)\in g}\left[(z_{nk} - P_g)^2 + (P_g - P_{nk})^2 - 2(z_{nk} - P_g)(P_g - P_{nk})\right] \end{aligned}$$

The first term of the last expression is BS(G) and the second term is the between-group variation of var$(P|G)$, so we need only demonstrate that the last term corresponds to the average covariance between z and P.

$$\begin{aligned} &\sum_{g=1}^{G}\sum_{(n,k)\in g}(z_{nk} - P_g)(P_g - P_{nk}) \\ &= \sum_{g=1}^{G}\sum_{(n,k)\in g}\left[(z_{nk} - z_g)(P_g - P_{nk}) + (z_g - P_g)(P_g - P_{nk})\right] \end{aligned}$$

Since $(P_g - P_{nk})$ sums to zero for each group g, while $z_g - P_g$ remains constant for these same groups, the second component of the last line is zero. The first component is the negative-covariance within each group g. ■

The better our model predictions at the individual level, the higher the value of the covariance of z and P. The more homogeneous the predictions within each group, the lower the variance of P within each group. With these definitions, in Table 2.2, we show the values of the Murphy Decomposition of the same set of predicted probabilities using

averages taken over several different groups defined by crossing a geographic zone and a choice of large-surface store.

Zone	Brier	BrierZ	VarByZ	CovByZ	UV	Resol	Calib
Cen	0.335	0.374	0.041	0.040	0.439	0.073	0.008
Dep	0.335	0.371	0.038	0.037	0.439	0.077	0.009
Sec	0.335	0.343	0.011	0.010	0.439	0.126	0.030
UU	0.335	0.342	0.008	0.008	0.439	0.149	0.052
Com	0.335	0.336	0.002	0.002	0.439	0.162	0.059
Iri	0.335	0.335	0.000	0.000	0.439	0.175	0.071
Ind	0.335	0.335	0.000	0.000	0.439	0.439	0.335

Table 2.2: Murphy Decomposition by geographic zone

"BrierZ" stands for the Brier Score where the predicted probabilities are replaced with the average predicted probabilities in the given geographic zone. "VarByZ" is the average variance of the predicted probabilities within each geographic zone, and "CovByZ" is the average covariance of the predited probabilities with the observed responses within each geographic zone. Here, the groups are ordered from largest to smallest. "Cen" represents a group containing all households in the region associated with each choice of alternative, "Dep" represents French departments, "Sec" represents survey sectors, "UU" represents INSEE-defined urban units, "Com" represents French communes, "IRI" represents "IRIS", and "Ind" represents individuals. Each geographic zone is in general an aggregation of smaller-sized zones[2]. As we can see, the resolution (the between-group variation) automatically increases as we go from larger to smaller groups. Moreover, we see that the variance of predicted probabilities increases as we go to larger groups, while this is closely related to the covariance of predicted probabilities and observed responses.

We see that from one well-known index, the Brier Score, we have created a great

[2]The only exceptions are some communes that contain multiple survey sectors, and some urban units not only contain several survey sectors, but can also span department boundaries

number of indices that we can choose according to our needs. If we care about the individual predictions, we would go with a Brier Score. If we are more concerned about aggregate predictions, we can define the geographic zone at which we do the aggregation, and then calculate a Calibration term. If we believe we wish to ignore the variation of predicted probabilities at an individual level, instead preferring average forecasts, we can look at predicted probabilites at any level of aggregation we wish, simply by taking the calibration term and discarding the variance and covariance of the residual terms.

2.2.3 Calibration as overlapping fluxes

We have overviewed some standard ways in which model fit is evaluated based on log likelihoods. However, our procedure in evaluating our model takes a slightly different approach. The way we validate our model is by simulating the use to which a forecaster would put this model; that is, assigning a prediction of store choices to each individual, and then evaluating the quality of such predictions.

Once we have determined our assignment strategy, we have two perspectives from which to calculate the accuracy of our assigned alternatives. The first is from the perspective of the individual. If a given person is assigned a given store, we would like to know how likely this prediction is wrong. In the second perspective, which we shall call the store's perspective, we are concerned with how close the predicted number of households in our sample choosing each store is to the actual number choosing each store.

We suppose that we have assigned a choice of alternative to each individual in our model, based on our predicted probabilities. This can be done by assigning the most likely alternative to each individual, or by drawing an alternative based on the predicted probabilities. To judge the accuracy of a model's individual predictions, we calculate the percent of individuals for whom the assigned alternatives corresponds to the observed

chosen alternative.

We use the following formula that we call "WA" (for "Well-Allocated"):

$$\begin{aligned} \text{WA(individual)} &= \frac{\sum_{n=1}^{N} \zeta_n}{N} \\ \zeta_n &= \begin{cases} 1, \text{ if } A_n = z_n \\ 0, \text{ otherwise} \end{cases} \\ z_n &= \text{Observed first choice} \\ A_n &= \text{First choice assigned} \end{aligned}$$

The number of false predictions at the individual level may not be an appropriate measure of the quality of our model. Suppose that two-thirds of the inhabitants of a city go to Store A, and one third go to Store B. If one model predicts that every inhabitant goes to Store A, then 66 percent of the households' choices of store will be well-predicted. Suppose a second model assigns a choice of store to each individual by making a draw from the individual's choice set using equal probabilities of selection so that half the individuals make choices that are well-predicted. If we look at the percentage of correct predictions, the first model is superior. However, the number of the city's households visiting each store is wildly inaccurate in the first model, while the second model has a number more resembling the real number of clients. Our preference for the second model of store choice depends on the importance of the identity of households making the purchases.

In order to have a criterion that favours the second model, we develop a different measure of model accuracy based on a comparison of the number of customers predicted for each store and the number observed. Suppose that in the example of the city described above, we consider households interchangeable. In our second model of store choice, store A is assigned to 50 percent of households, but is chosen by 67 percent, and Store B is

assigned to 50 percent of households but is chosen by 33 percent. Since there are less households observed selecting Store A than households to whom this store choice is assigned, we can reassign store choices, swapping assigned choices of Store A amongst households observed not selecting Store A and assigned choices of Store B amongst households observed selecting Store A, so that every household observed selecting Store A will have the correct choice of store assigned to it. Doing this, we can have 83 percent of households in the population that are well-assigned, of which 50 percent of the population corresponds to households observed selecting Store A and now assigned Store A, and 33 percent of the population corresponds to households observed selecting Store B and now assigned the choice of Store B. The maximum number of households that can be well-assigned doing this will thus be the sum of the minima of the number of households assigned to each store, and the number of households observed selecting each store.

Our measure of model fit therefore comes down to determining a geographic zone over which we assume that households are interchangeable, and then a calculation of the overlap between the number of households predicted visiting each large-surface store and the number observed. If we define by O_{sk} and A_{sk} the number of households observed living in zone s and selecting alternative k, and the predicted number of households living in zone s who choose the alternative k, then this overlap can be measured by the following metric that we call "WD" (for "Well-Distributed"):

$$\text{WD(Geographic Zone)} \quad = \quad \frac{1}{N} \sum_{s=1}^{S} \sum_{k \in C_s} \min(O_{sk}, A_{sk})$$

"WD" is a number between zero and one, one being the case where the number of households in each zone observed selecting each large-surface store equals the number of households predicted visiting each large-surface store. A lower value of "WD" indicates less overlap between predicted and observed numbers of households from each zone vis-

iting each store, and therefore, in general, a greater difference between predicted and observed numbers of store visits.

We note that the value of A_{sk} can be calculated following any way of assigning alternatives. This can be

$$A_{sk} = \sum_{n \in s} \sum_{k \in C_s} A_{nk}$$

where s is the set of all households in zone s, and A_{nk} is one when alternative k is assigned to household n either by being the alternative with the greatest predicted probability, or is the alternative that is assigned to n through a random draw weighted by the predicted probabilities of selecting each alternative in n's choice set. Or, A_{sk} can be equal to

$$A_{sk} = \sum_{n \in s} \sum_{k \in C_s} P_{nk}$$

This function, is, in fact, a function of the sum of the absolute values of the residuals of the aggregated model predictions by zone. We see this in the fact that

$$\begin{aligned}
\frac{1}{N} \sum_{s=1}^{S} \sum_{k \in C_s} \min(O_{sk}, A_{sk}) &= \frac{1}{N} \sum_{s=1}^{S} \sum_{k \in C_s} \frac{O_{sk} + A_{sk} - |O_{sk} - A_{sk}|}{2} \\
&= \frac{1}{N} \left(\frac{N}{2} + \frac{N}{2} - \sum_{s=1}^{S} \sum_{k \in C_s} \frac{|O_{sk} - A_{sk}|}{2} \right) \\
&= 1 - \frac{1}{2N} \sum_{s=1}^{S} \sum_{k \in C_s} |O_{sk} - A_{sk}|
\end{aligned}$$

By taking s as a geographic subdivision g, and A_{sk} the average predicted probability of households in s selecting k, and O_g, the average number of households in s selecting k, we see that our formula contains the calibration from the Murphy Decomposition of

the Brier Score, only with absolute values replacing squared terms:

$$\text{Alternative calibration} \quad = \quad -\frac{1}{2N}\sum_{g\in G}\sum_{(n,k)\in g}|z_g - P_g|$$

We have used an example to illustrate how the calculation of our procedure works. Figure 2.1 shows the survey sectors and the communes in the Agglomeration of Tours.

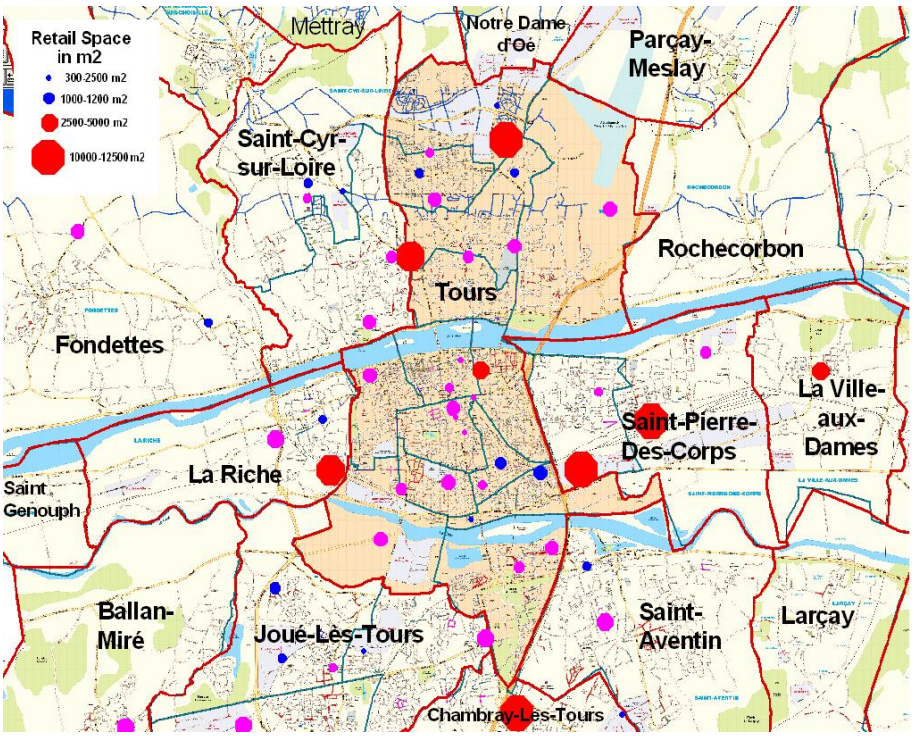

Figure 2.1: Map of the city of Tours indicating divisions into communes (red lines) and survey sectors (blue lines) and indicating hypermarkets large and small (red), supermarkets (purple) and hard discount stores (blue).

In Figure 2.2 We look at the three survey sectors at the centre and bottom of the map in Figure 2.1, and look at the three closest large hypermarkets to it. We name the three sectors, Sectors 1, 2, and 3, and the three hypermarkets, Store A, Store B, and Store C.

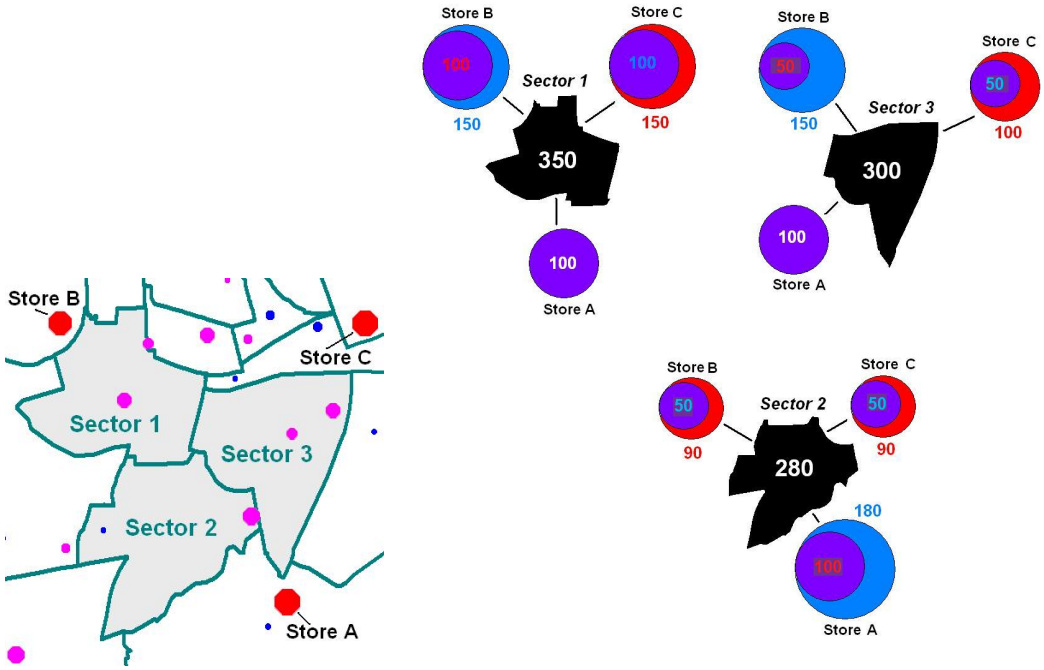

Figure 2.2: Illustration of "WD" statistic calculated on three survey sectors in Tours.

In Figure 2.2, we illustrate the calculation of this statistic with a hypothetical example. Suppose we have three sectors (Sectors 1, 2, and 3) and we have three stores (Stores A, B, and C). We count the number of households in each Sector (350, 300, and 280, respectively) and count how many are observed to go in each store, and compare this with the number predicted going into each store. The red circles, and the numbers in red, represent the number of households observed going into each store. The number of households that our model predicts would go in each store from each survey sector are represented by the blue circles, and the numbers in blue. The circles in purple represent the maximum number of households who can be assigned the large-surface store it is observed to visit, if the total number of households assigned to each store is equal to the total number predicted by the model. When there are the same number of observed as predicted choices, there is a single purple circle with a number in white representing the number of observed and predicted choices. The WD statistic for this small example will therefore be the sum of the number of households within the sets of overlapping predicted and observed store choices, and the total number of households, or, the sum of the numbers within the circles divided by the sum of the numbers within the outlines of the survey sectors:

$$\begin{aligned} \text{WD} &= \frac{100+100+100+50+50+100+50+50+100}{350+300+280} \\ &= \frac{700}{930} = 0.753 \end{aligned}$$

We can thus produce a table comparing the different results, using this new statistic. Table 2.3 shows the Brier Score, the Brier zone calculated with averages of predicted probabilities by zone, the Calibration defined using squares of differences between observed and predicted shopping flows, the calibration defined using the sums of absolute differences between observed and predicted shopping flows, and then the "WD" statistic.

Zone	Brier	BrierZ	Calib	CalibAbs	WD
Cen	0.335	0.374	0.008	0.096	0.904
Dep	0.335	0.371	0.009	0.109	0.891
Sec	0.335	0.343	0.030	0.217	0.783
UU	0.335	0.342	0.052	0.246	0.754
Com	0.335	0.336	0.059	0.272	0.728
Iri	0.335	0.335	0.071	0.320	0.680
Ind	0.335	0.335	0.335	0.668	0.332

Table 2.3: Calibration and Brier Score by geographic zone

Looking at these closely-related terms, we feel that the "WD", is the most meaningful. It is near one for the Centre Region, since the estimates of the number of people visiting each store over the entire region is very close to the observed number, and it is closer to zero at the individual level, since there is a great divergence between predicted and average probabilities since our model has trouble distinguishing between individuals.

We recall that our "WD" statistic validates our model by comparing predicted probabilities with observed store choices. In Table 3.16, the model used to assign predicted probabilities is based on the same observations whose corresponding observed store choice is used to validate these predictions. This method risks allowing overfit to remain undetected, and is not a good judge of the accuracy of a model's predictions on other samples.

We therefore use the technique of cross-validation to test the accuracy of our model predictions. In order to do this, we divide our set of households into subsamples according to the department of residence of each household. The predicted probabilities of selection assigned will be done separately for each department. When assigning the probabilities of selection to one department, we calculate the estimated parameters of the model using a "training set" of observed store choices of all households not living in the department in question, and then use these parameters to calculate the predicted probabilities of

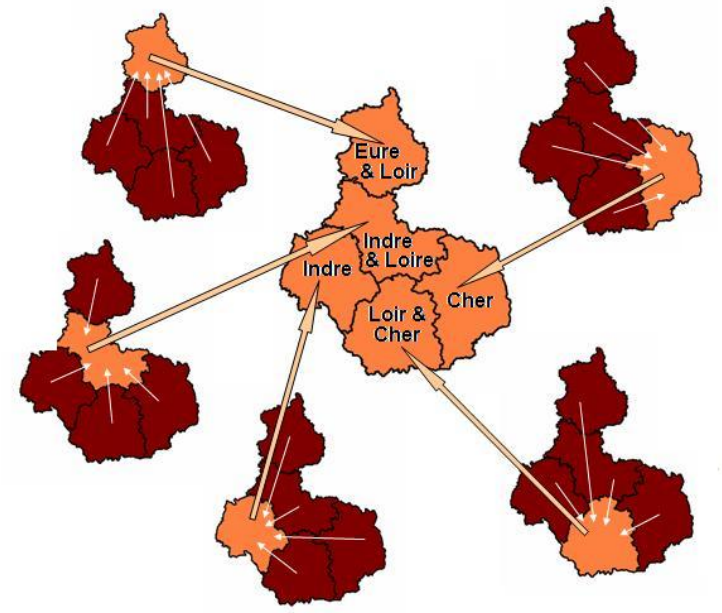

Figure 2.3: Cross validation of imputation by department.

selection for the "test set" of households living in the department. For example, if we were to predict the probabilities of selection of each store for households living in the Indre department, we would calculate the estimated parameters of our model by running our estimation procedure on a data set containing the choices of large-surface stores made by households in the Eure-et-Loir, Indre-et-Loire, Loir-et-Cher, and Cher departments. Once probabilities of selection have been assigned to each department, we can combine these predictions into one data set containing probabilities of selection for every single household in the survey region. This will work as a simulation of the use of our model using parameters calculated from the households within the survey region on a different survey region that is independent of the original sample, but resembles the original survey region. We illustrate this in Figure 2.3, where we show in dark the areas in the survey region that are used to form the training sets, and the light areas represent the test sets.

2.3 Summary

In this chapter, we looked at both the issues of how to define the choice set for the prediction of the choices of large-surface store made by each household, and how to evaluate these predictions. In Section 2.1, we described how we could not conduct Conditional Logit estimation on a data set in which choice sets contained all the possible choices of large-surface stores in the sample. In fact, there was a difference between the universal set of large-surface stores that could be chosen by the household, and the consideration set of large-surface stores that the household would choose from. We considered a few solutions. The model proposed by Basar and Bhat (2004 [5]) improved the model estimation by supposing that the choice set of the individual was generated by random effects, but it took longer to calculate and could not be applied to our case. Another solution was to consider only the closest stores to the household and consider all other store choices as comprising the alternative labelled the "outside option". This could also serve as a more realistic supposition of the choice set of each household. A third possibility was to draw alternatives from the universal choice set at random, and then use the estimated parameters in order to assign probabilities of selection to all households crossed with all large-surface stores. It has been shown by McFadden (1978 [34]) that using sampled alternatives results in maximum likelihood estimates of the parameters that are consistent with estimation done over the universal choice set, however, for this to be true, the sampling strategy would need to meet the "Positive Conditioning Property". If the sampling strategy does not meet a further condition known as the "Uniform Conditioning Property", then one would need to enter an adjustment term into the maximum likelihood estimate based on the probabilities of drawing the sample in order to maintain consistency. There are different sampling strategies that meet the Positive Conditioning Property: simple random sampling, probability proportional to

size sampling drawing with replacement, and probability proportional to size sampling drawing without replacement. Only in the simple random sampling is the Uniform Random Sampling Property verified. Since the probability of drawing a choice set is extremely difficult to calculate for sampling without replacement, we prefer sampling with replacement when we define a choice set.

In order to measure model quality, we can look either at the fit of the model's predictions to the data set on which it is estimated, or we can measure the accuracy of the model's predictions when compared with the observed choices of stores. One measure of the former is the McFadden Pseudo R-squared value, and one measure of the latter is the Brier Score. Since we know the uses to which our models will be put, we develop our own criterion for model evaluation. Since what interests BVA is the accurate prediction of shopping flows from residential zones to large-surface stores, we develop what we call the "WD" statistic based on the amount of overlap between the forecast and observed flux between residential zones and large-surface stores. It turns out that comparing the Brier Score for different models applied to the same data is equivalent to comparing the Calibration of the different models. This Calibration is very similar to the "WD" statistic that we developed. In order to account for the effect of applying our model to a data set other than the one that was used in order to estimate the model parameters, we calculate the "WD" statistic with cross-validation by department of residence of each household.

Chapter 3

Application of Store Choice Models to Survey Data: Estimation and Comparison of Results

In this chapter, we apply the modelling procedures that we developed in Chapters 2 and 3 to our survey data and evaluate the results. We are interested in doing this in order to compare and evaluate three different techniques of assigning probabilities to store choices: a Conditional Logit model of store choice, a "gravitational model", and a "Hybrid" model of store choice that combines a Conditional Logit model with a gravitational model. Each of these three techniques will be described in a different subsection. In Section 3.1, we describe how we estimate the Conditional Logit model on our data, which, since our computations are not feasible when every possible choice of store is treated as a possible alternative, we do using both aggregated and sampled alternatives. We then describe the variable selection techniques that we used to determine the set of explanatory variables that we enter into our Logit expression. We conclude

by presenting a comparison of the predictive accuracy of the different estimations. In Section 3.2, we describe our use of a gravitational model of store choice based on the early market-share models that predated the Logit modelling technique and in which we use a simple regression to determine the model parameters. We present this procedure, comparing the accuracy of different specifications of the model. Section 3.3 refers to the use of Logit estimation in order to assign probabilities to aggregated models, and then the use of the gravitational model in order to assign probabilities to rare store selections. This allows the assignment of store probabilities to all stores within the data set while improving the predicted probabilities of the most common choices of large-surface stores. The chapter ends with a final evaluation of the different models used, and looks at the possible extensions of the model.

Before we describe in more detail the techniques that we test here, we would like to discuss the general criteria by which we evaluate our store predictions for every type of choice probability assignment. We defined formal criteria in Section 2.2 for selecting the model that produces the most reliable predictions, but since we are working for a private survey company, the most important question we need to answer is whether we can possibly create a product that can be sold at a profit. This means that we need to take into account the monetary cost of developing and using each model in our assessment of each model. When we propose different techniques to BVA, then, we look at the following criteria:

- Computational burden
- Ease of implementation
- Reliability of predictions

Our model assessment must remain very sensitive to the first two criteria, since they represent the cost to BVA of the model's use. Since the model will be used on very

large data sets in order to generate predictions on an enormous number of individuals, our calculations face much greater time and resource constraints than are found in many academic papers dealing with discrete choice models, which typically make use of smaller data sets. Because BVA intends to continue using the model after it has been developed, we must take into consideration the ease with which other employees can be trained to use the model. The cost to BVA can also be divided into fixed costs and variable costs. The fixed cost involves the cost of adapting (or training) the model to the survey data we have been provided, and the variable cost involve the cost of using our model to calculate predictions of shopping behaviour. Computational burden and user friendliness will be aspects of both these types of costs. A product's reliability is what will eventually "sell the product." These considerations must be borne in mind as we examine every type of modelling technique.

3.1 Conditional logit model estimation

We have described the Logit model extensively in earlier sections. Here, we shall compare eight different Conditional Logit models that we tested on our survey data. The first four models have choice sets developed using the method described in Section 2.1.2 in which we aggregate alternatives with low probabilities of selection. The last four are models in which choice sets were developed through the sampling of available alternatives.

Conditional Logit with aggregated choice sets:

We define a choice set containing as alternatives the supermarkets, hypermarkets, and hard discount stores that are closest to each household's home, in addition to which there is one alternative comprising all other stores, and one alternative consisting of no choices of large-surface stores. The four choice set definitions we created are labelled

Model	Choice set	# alt	% outside
7694	supermarket within 7 closest to home, small hypermarket within 6 closest to home, hard discount within 9 closest to home, large hypermarket within 4 closest to home, no store or other stores	25+	1^{st} choice: 8 % 2^{st} choice: 7 % 3^{st} choice: 3 %
4372	supermarket within 4 closest to home, small hypermarket within 3 closest to home, hard discount within 7 closest to home, large hypermarket within 2 closest to home, no store or other stores	18+	1^{st} choice: 13 % 2^{st} choice: 12 % 3^{st} choice: 6 %
3232	supermarket within 3 closest to home, small hypermarket within 2 closest to home, hard discount within 3 closest to home, large hypermarket within 2 closest to home, no store or other stores	12+	1^{st} choice: 18 % 2^{st} choice: 18 % 3^{st} choice: 9 %
2121	supermarket within 2 closest to home, small hypermarket within 1 closest to home, hard discount within 2 closest to home, large hypermarket within 1 closest to home, no store or other stores	8+	1^{st} choice: 31 % 2^{st} choice: 31 % 3^{st} choice: 15 %

Table 3.1: Choice sets of aggregated alternatives

"7694", "4372", "3232", and "2121" and are described in Table 3.1. We must emphasize that with these choice set definitions, we cannot use our model to predict the probability of any one household visiting any large-surface stores. This loss of information must be taken into account when we compare this technique to others.

In Table 3.1, we have included in each column the number of alternatives in each choice set (which can be slightly greater if there are tied distances, for example in the case where there are three stores that are tied as being the closest stores to a households' domicile there will be three stores counted as being within the two closest to the household's home.) We have also included the percent of households choosing the outside option associated with each choice set for each of their three choices of large-surface stores.

Conditional Logit with sampled choice sets:

In addition to aggregating alternatives to create an "outside option", we have defined the choice sets for our model by sampling alternatives, following the techniques described in Section 2.1.3. These models are shown in Table 3.2 In the first two models, we draw a sample of alternatives using Probability Proportional to Size (PPS) random sampling, in which the sampling probabilities are determined by a function of the distance of the store and the store type. The function we use to determine the weights of this sample is in fact the utility function of the modified gravitational model that we describe in Section 3.2. We found that with 18 random draws for each individual, after the addition of the chosen alternative we would have on average about 12 alternatives per household, the same size of choice set as the one found in the "3232" model in the aggregated choice models.

Here "Outside" would be one for all stores that are in the outside option of the "3232" model. Similarly, the choice set of the second model is created through 31 random draws, that, with the addition of the chosen alternative, correspond to an average of 18 alternatives per household, just like in the "4372" model. Here, we define the outside option as being the alternative that is considered within the "outside option" of the "4372" model.

To contrast these two models with choice sets drawn using PPS random sampling, we do the same, only using simple random sampling (SRS), with the same number of draws, and the same definitions of "outside". In SRS, every draw yields a unique selection, so the number of alternative in each choice set will be equal to roughly the number of draws undertaken.

Model	Choice set	# alt	Sample
R18	18 random draws of alternatives from all possible store choices (including chosen stores), with probabilities weighted by distance, and the observed store choice added to the choice sets	12	PPS ($\approx$ 3232)
R31	31 random draws of alternatives from all possible store choices (including chosen stores), with probabilities weighted by distance, and the observed store choice added to the choice sets	18	PPS ($\approx$ 4372)
NG11	11 options are drawn at random from all possible store choices (including chosen stores), with equal probability of selection, and the observed store choice is added to the choice sets	12	SRS ($\approx$ 3232)
NG17	17 options are drawn at random from all possible store choices (including chosen stores), with equal probability of selection, and the observed store choice is added to the choice sets	18	SRS ($\approx$ 7694)

Table 3.2: Choice sets of sampled alternatives

3.1.1 Selecting explanatory variables in a Conditional Logit model

When we do Conditional Logit estimation, whether we aggregate or sample alternatives, we must determine the set of explanatory variables upon which our model estimation will be based. We have have at our disposal a great deal of information on the households in our survey sample and the large-surface stores in their region, but we cannot use all the variables in our data sets as explanatory variables in our model, since they contain a great deal of redundant information. There is a lot of collinearity between many of our variables, and many variables have no explanatory value for the given store choice. Once we have specified our choice set, therefore, we need to choose which variables we will include in the design matrix X that will be entered into the probability defined by the Logit Model:

$$P_{nj} = \frac{e^{\beta X_{nj}}}{\sum_{k \in C_n} e^{\beta X_{nk}}}$$

This must be done with both field expertise and technical analysis. We first need to determine whether there is any collinearity in the set of variables entered into the X matrix. If some variables are linear combinations of others, then the X matrix will not have full column rank, rendering Newton-Raphson estimation impossible, since the X^TX matrix will not be invertible. If variables are close to collinear, parameter estimates become erratic as the determinate of X^TX approaches zero. For this reason, we need to eliminate any variables that we know to be collinear, or nearly collinear. Fortunately, the procedure that we use to estimate our model (the MDC or "Multinomial Discrete Choice" Procedure in SAS) is equipped to detect variable collinearity, something that can help us in cases where it is not as easy to see. Once we have determined a set of variables that are linearly independent, we then need to find the subset of these variables that provides optimal fit. The number of explanatory variables in our X matrix determines the degrees of freedom of our model. In general, if we have too many degrees of freedom, then we will be introducing variation into our model that is unrelated to the variation of the variable of interest, thereby creating less accurate predictions. If we have too few degrees of freedom, then we are not including some effects that play a role in our households' behaviour, and we risk biasing our estimators. Some econometricians will favour inaccuracy over bias, since the error due to bias is harder to quantify.

The way we select the variables that we will retain in our model is by testing the significance of all model parameters using a Wald statistic, that is, the ratio of the parameter estimate divided by their estimated variances, as determined by the Hessian matrix. The null hypothesis to be tested is that β_c, a component of the parameter vector β is zero. The alternative hypothesis is that it is not zero:

$$H_0 : \beta_c = 0 \quad H_A : \beta_c \neq 0$$

Since $\frac{\hat{\beta}_c}{\hat{\text{var}}(\hat{\beta}_c)}$ is a Wald statistic (with $\hat{\text{var}}(\hat{\beta}_c)$ taken from the cth diagonal entry of the Hessian matrix), we can calculate a p-value for the event that β_c is not equal to zero by assuming that this statistic follows a student-t distribution. The importance of this test is in evaluating whether to consider the estimated parameters in our model as representing real effects, or whether they are simply the result of random factors particular to the data set in question.

In order to select our variables, we develop a method of automatic variable selection. We proceed in several steps. At each stage, we calculate the Wald statistics of all parameters, and then eliminate the parameter with the lowest p-value, if it is below 0.15. In most statistical tests, the default level of significance is at a p-value of 0.05. However, due to the great importance of the inclusion of all significant parameters in a model, we have used a greater p-value to reduce the possibility of biassing the sample through omitting variables.

We cannot eliminate all nonsignificance parameters in one step, since the Wald statistic is a test of the significance of a single parameter within the model, and not the significance of a group of parameters. In fact, some parameters may be significant when evaluated individually, but not when evaluated as a group, and some parameters may be significant at the beginning of our process, but significant when other variables are removed. We can carry on this process until we arrive at a stage where every parameter in our model has a p-value less than 0.15, and therefore we consider significant.

Our ultimate goal is not so much identifying the causes of consumer behaviour (although that question is very interesting to us), but in predicting it, and this affects the way we look at these t-tests. Our use of these tests is in order to judge whether a model parameter estimated using one data set will be valid in a model used to predict the behaviour of individuals in another data set.

The estimates of the parameters calculated using SAS give us a model that is fitted to a particular sample of individuals that we shall call a training set. If we wish to use the same model in order to represent the behaviour of individuals in another sample, we would need to recalculate the maximum likelihood estimates of the model parameters using the data from the new sample. Unfortunately, we cannot calculate parameter estimates that are adapted to a set of individuals (called the prediction set) whose behaviour we wish to predict, since their behaviour is unknown. We therefore set as the parameters of the model that we use on the prediction set the estimates of the parameters of the model on a training set. The validity of this method will depend upon the degree to which the values we set for the model parameters would have been different had the behaviour of the individuals in the prediction set been known and used in order to generate the model parameter estimates. This will depend first of all on whether we can be assured that the individuals in the prediction and training data sets follow the same data generating process. This is not something that can be read directly from our data, and so we must rely on our judgement. In our case, our initial confidence is based on BVA's expertise that there is enough stability in French supermarket choices to justify using a model based on one region of France to make predictions for the entire country. Our confidence is enhanced by a study (Severin, Louviere and Finn, 2001 [40]) that shows that the maximum likelihood estimates of the parameters of conditional logit models of supermarket choices remained stable when applied to different countries and to different time periods.

Unfortunately, even if we assume that the individuals in the training and prediction sets follow the same patterns of behaviour, as we do, we need to be assured that if maximum likelihood estimations were done on both data sets that random effects wouldn't cause the parameter estimates to differ. This is where the p-values of the parameter

estimates are very useful.

The p-value is the probability that if we take as a null hypothesis that the true value of the model parameter be zero, and this null hypothesis holds for our model, that this hypothesis will be rejected if we take the estimated parameter in our model as being significantly different from zero. In other words, the p-value is the probability that our maximum likelihood estimates a value for the given parameter is further from zero than the parameter estimate, if the true model parameter is zero. However, if we recentered our t-statistic, it is also the probability that if the true model parameter were equal to the value we estimated, that the estimated parameter could be less than zero, or twice as large. Since the test statistic has a symmetric distribution, the probability that the estimate could be less than zero given that the true parameter is equal to the one we estimated, is simply half the p-value.

This means that for example, if the p-value of the estimated coefficient of a given variable in our model is 0.24, and this estimated coefficient is in fact exactly the true model parameter for the data generating process producing both the training set and all prediction sets, there will be a probability of 0.12 that the coefficient of this variable best adapted to a prediction data set of the same size as the training set will in fact be negative and the relationship between effect and behaviour will be reversed. Thus, in order to ensure that our model's coefficients won't "flip" in this way when we use the model to predict probabilities of selection for individuals not included in the training set, we take care to choose a set of parameters that not only have intuitive interpretations, but that have low p-values.

We must take care in eliminating nonsignificant variables. We cannot simply eliminate all variables with high p-values, since these depend on the other variables included in the model. The order in which we eliminate variables may also determine which vari-

ables we end up with when we have only significant variables left. If we are left with a model with significant effects but that go against our understanding of the behaviour represented by our data, we can attribute this to the limitations of our model and can begin our process of data selection again, eliminating variables not only with high p-values, but with signs that are contrary to our expectations. We must remember that there may not be a unique set of variables that reflects the effects present in our model. Our challenge is to select the set of variables that lend themselves best to a logical interpretation.

If we introduce the variables that characterize the individual decision-maker, but are constant with respect to different alternatives, they would be entered as cross-terms of these variables and the explanatory variables dependent on alternative characteristics. When we do this, we find that the socio-demographic variables are not significant, and in a model in which there is already an enormous number of explanatory variables, this causes the parameter estimates of the model to become complex beyond possible interpretation.

In Tables 3.3 and 3.4 we list all the explanatory variables that we include in our model. Binary dummy variables are constructed to represent all multinomial variables. In Table 3.3 we list the choice-specific constants of our model. A model containing only these variables will generate probabilities of selection of each alternative roughly equal to the percent of households in the sample selecting each alternative. (The probabilities are not exactly equal to market shares since tied distances render choice set sizes unequal). To create variables of type "SMRankGE2", we calculate the distances between the coordinates of each household's home (which correspond to the centroid of the IRIS in which the household lived) and every single store in the population. We then rank all stores of the same type in ascending order according to distance. Tied distances are low, meaning that, for example, two stores tied for second closest are assigned rank 2,

while the next closest store is assigned rank 4. Thus, the variable "SMRankGE3" is one if the alternative in question corresponds to a supermarket and there are exactly 2 supermarkets whose co-ordinates are nearer the household's home co-ordinates (the centroid of its IRIS of residence.) Variables of type "OutWHDNumGE12" adjust the utility of the "outside" option with respect to the number of alternatives included in each choice set. Thus, for example, "OutWHDNumGE12" is one when the alternative in question is the "outside" option and there are at least 12 hard discount stores included as separate options within the household's choice set and zero otherwise.

In Table 3.3, variables are all indications of the characteristics of each large-surface store in a household's choice set and the geographic co-ordinates of its location crossed with the store type. For example, disSM is the product of dis and SM and represents the euclidean distance of a supermarket from the household's domicile in km. Similarly, SMGStu299le0 is also the product of GStu299le0 and SM. Most of these variables are based on public information on the characteristics of French communes. One can refer to Section 1.1.4 for the definition of the polarity of communes.

Probability forecasting is too costly when we make use of the entire choice set of individuals, so to render the forecasting more efficient, if a drawn store were found in the outside option of the "3232" model, then we set all the following explanatory variables to zero except:

- SM, HM, HD, XM
- disSM, disHM, disHD, disXM
- dissqSM, dissqHM, dissqHD, dissqXM
- DensPopuSM, DensPopuHM, DensPopuHD, DensPopuXM
- AccAutoRSM, AccAutoRHM, AccAutoRHD, AccAutoRXM
- AccAutoRZSM, AccAutoRZHM, AccAutoRZHD, AccAutoRZXM

Variable	Description	Type
SM	Supermarket	Dich
SMRankGE2	Supermarket with rank of distance >= 2	Dich
SMRankGE3	Supermarket with rank of distance >= 3	Dich
HM	Small hypermarket	Dich
HMRankGE2	Small hypermarket with rank of distance >= 2	Dich
HD	Hard discount store	Dich
HDRankGE2	Hard discount store with rank of distance >= 2	Dich
HDRankGE3	Hard discount with rank of distance >= 3	Dich
XM	Large hypermarket	Dich
XMRankGE2	Large hypermarket with rank of distance >= 2	Dich
outside	Outside option chosen ("other stores")	Dich
OutWSMNumGE3	Outside option for choice set with 3 or more alternatives representing supermarkets	Dich
...	...	...
OutWSMNumGE12	Outside option for choice set with 12 or more alternatives representing supermarkets	Dich
OutWHMNumGE3	Outside option for choice set with 3 or more alternatives representing small hypermarkets	Dich
...	...	...
OutWHMNumGE12	Outside option for choice set with 12 or more alternatives representing small hypermarkets	Dich
OutWHDNumGE3	Outside option for choice set with 3 or more alternatives representing hard discount stores	Dich
...	...	...
OutWHDNumGE12	Outside option for choice set with 12 or more alternatives representing hard discount stores	Dich
OutWXMNumGE3	Outside option for choice set with 3 or more alternatives representing large hypermarkets	Dich
...	...	...
OutWXMNumGE12	Outside option for choice set with 12 or more alternatives representing large hypermarkets	Dich
Nostore	No store chosen	Dich

Table 3.3: Basic Variables

Variable	Description	Type
dis	Euclidean distance of store from home in km	Cont
dissq	Square of dis	Cont
surf	Surface area of supermarket in thousands of m^2	Cont
surfsq	Square of surf	Cont
Samedep	Large-surface store is in same department as household's residence	Dich
SameUU	Large-surface store is in same commune as household's residence	Dich
Samecit	Large-surface store is in same commune as household's residence	Dich
GStu299le0	Commune of large-surface store classed as rural	Dich
GStu299le1	Store is in nonrural commune with population less than 10K inhabitants	Dich
GStu299le2	Store is in nonrural commune with population less than 50K inhabitants	Dich
GStu299le3	Store is in nonrural commune with population less than 100K inhabitants	Dich
GSpol99le1	Commune of large-surface store classed as urban pole	Dich
GSpol99le2	Commune of large-surface store classed as urban pole or monopolarized	Dich
GSpol99le3	Commune of large-surface store classed as urban pole, monopolarized, or multipolarized	Dich
GSVC99_1	Commune of large-surface store classed as city centre	Dich
TR2ROU	Percent of population in household's home commune commuting to commune of store in question by a two-wheeled vehicle	Cont
TRCOM	Percent of population in household's home commune commuting to commune of store in question by public transportation	Cont
FavCom	Store's commune is the most visited by those in commune of household's domicile	Dich
FavVil	Store's commune is the most visited by those in commune of household's domicile and it has more than 10,000 residents	Dich
FavCom	Commune of store is the commune most-visited by population living in commune of household's home	Dich
AccComF	Time in hours to go from domicile to commune of store in question if it is the most frequently visited commune by those living in commune of household	Cont
AccVilF	Time in hours to go from domicile to commune in question if it is the most frequently visited commune by those living in commune of household and it has more than 10,000 residents	Cont
AccAutoR	Time in minutes to access closest autoroute	Cont
AccComFZ	AccComF is zero or missing	Dich
AccVilFZ	AccVilF is zero or missing	Dich
AccAutoRZ	AccAutoR is zero or missing	Dich
Denspopu	Population density of commune of large-surface store	Cont

Table 3.4: Variables Crossed with Store Type

- Outside

3.1.2 Conditional Logit Parameter Estimates

In Tables 3.5, 3.7, 3.8, 3.9, 3.10, and 3.11 we present the parameter estimates yielded by each of the eight models we tested for the first choice of large-surface store. In Table 3.5, we compare all the estimates of the parameters representing store-specific constants in each model. In each of the tables, we highlight in blue the variables whose estimated coefficients have different signs in different models.

In general, looking at these tables, we can say that parameter estimates remain remarkably stable as we go from one choice set definition to another, although the models differ in the parameter estimates that are determined to be nonsignificant. The signs of the parameter estimates remain the same for all the initiated models, except only a few cases that are marked in blue. This happens rarely enough that we can attribute this to random error and imperfect model specificiation.

In Table 3.5, we look at the coefficients of supermarket-related dummy variables associated with each model. Some of the differences between the estimates of each model are clearly dependent on the different choice set definitions. The variable "SMRankGE3" indicates that the alternative in question is a supermarket that is further from the given household than at least two other supermarkets. Clearly this variable is not relevant in the model "2121", since no alternative corresponds to such an alternative in that model. The fact that all variables of the style "SMRankGE3", "HMRankGE2", etc are negative simply means that even taking into account a store's distance, and the many other components of utility in our model, the probability of a household selecting a given large-surface store will decline if there are more large-surface stores that are closer to the household than the store in question.

Parameter	2121	3232	4372	7694	NG11	NG17	R18	R31
SM	1.48	.	.	.	-1.83	-1.67	-0.74	-0.63
SMRankGE2	-0.34	-0.29	-0.29	-0.27	-0.39	-0.45	-0.44	-0.56
SMRankGE3	.	-0.43	-0.43	-0.42	-0.82	-0.80	-0.42	-0.51
SMRankGE4	.	.	-0.42	-0.42	.	-0.40	.	-0.43
SMRankGE5	.	.	.	-0.37	.	.	.	.
SMRankGE6	.	.	.	-0.25	.	.	.	.
SMRankGE7	.	.	.	-0.29	.	.	.	.
HM	.	.	.	.	-1.69	-0.84	-0.53	0.43
HMRankGE2	.	-0.18	-0.22	-0.20	.	-0.74	-0.58	-0.70
HMRankGE3	.	.	-0.64	-0.68	.	-0.86	.	-0.84
HMRankGE4	.	.	.	-0.81	.	.	.	.
HD	.	.	.	.	-1.18	-1.28	-2.11	-2.94
HDRankGE2	-0.22	-0.26	-0.28	-0.28	-1.23	-0.96	.	.
HDRankGE3	.	.	.	.	.	.	0.20	.
HDRankGE4	.	.	-0.64	-0.75	.	-0.59	.	-0.38
HDRankGE5	.	.	-0.31	.	.	-0.62	.	.
HDRankGE7	.	.	.	.	.	.	.	0.52
HDRankGE8	.	.	.	-0.80	.	.	.	.
XM	.	.	.	.	2.53	2.13	3.52	3.44
XMRankGE2	.	-0.96	-0.93	-0.93	-0.83	-0.58	-0.96	-1.12
XMRankGE3	.	.	.	-0.57	.	.	.	.
XMRankGE4	.	.	.	-0.26	.	.	.	.
outside	1.16	-1.60	.	-1.15	.	0.64	-3.44	-2.17
OutWSMnumGE8	.	.	.	-0.83	.	.	.	.
OutWSMnumGE9	.	.	.	0.84	.	.	.	.
OutWHDnumGE10	.	.	.	0.28	.	.	.	.
nostore	-1.77	-3.99	-2.04	-2.75	.	.	.	.

Table 3.5: Parameter Estimates Part 1: store-specific constants

The variables of the style "OutWSMNumGE8" are introduced in order to take into account the fact that there are many tied distances between stores and household domiciles so that there could be a variable number of large-surface stores that are identified as separate alternatives, something that could affect the probability of selecting the alternative representing outside stores. For example, in the "2121" model, if there were no ties for one household, it would face 8 alternatives, including the "outside" alternative that would be defined as "any store but the two closest supermarkets, the closest small hypermarket, the two closest hard discounts, and the closest large hypermarket." If there were two supermarkets that were tied as being the second closest to the household, both these stores would be classed as separate alternatives in the model, and the outside option would now be defined as "any store but the *three* closest supermarkets, the closest small hypermarket, the two closest hard discounts, and the closest large hypermarket". In the second case, the outside option would contain less stores, and could therefore be regarded as less probable than it would be in the first case. For this reason, we introduce the adjustment terms "OutWSMGE8", "OutWSMGE9" and so on so that the utility of the outside option varies with respect to the number of large-surface stores of each type included as separate alternatives in the households' choice sets. These adjustment terms only appear in the "7694" model where there is enough variation in the choice sets for these effects to be significant. "OutWSMGE8" is one when the alternative in question is the outside option, and there are at least 8 supermarkets that are considered separate alternatives in the choice set. In other words, this parameter is the difference in the utility of the outside option of a choice set containing at least 8 supermarkets treated as separate alternatives in the choice set, and less than 8 supermarkets treated as separate alternatives. In Table 3.6, we look at how many households correspond to each possible number of stores of each type in their choice set. For example, we see in the table,

	Number of households			
Store Type	SM	HM	HD	XM
4 Stores	0	0	0	12573
5 Stores	0	0	0	0
6 Stores	0	12338	11786	0
7 Stores	11331	235	724	0
8 Stores	1111	0	46	0
9 Stores	83	0	8	0
10 Stores	22	0	9	0
11 Stores	7	0	0	0
12 Stores	19	0	0	0

Table 3.6: Distribution of households by number of stores of each type in their choice sets.

that out of 12,573 households, 11331 had 7 stores in their choice sets, while only 19 had exactly 12. We also see that for every household without exception, there were 4 large hypermarkets defined as separate alternatives in their choice set.

In Tables 3.7 to 3.10, we look at all the variables specific to each type of large-surface store. We need to cross all the variables characterizing each type of large-surface store with the store type, since the different store types exhibit completely different relationships between store characteristics and utility. We see that in every model, the estimated coefficient of distance ("disSM","disHM", etc.) is negative, and the estimated coefficient of retail space ("surfSM","surfHM") is positive, since households are more likely to shop in stores that are large or are near than are far or small. We also see that the more accessible a commune is to a household, the more likely it will select a store that is there. For this reason, "samedepSM", "samecitSM", "TR2RouSM", "TRComSM", "FavComSM", and "FavVilSM" are all positive and "AccAutoR" is negative.

It may be surprising that "AccFavComSM" and "AccFavVilSM" are positive, since these are all indicators of the accessibility of the commune of the store in question. If "FavComSM" is positive, it means that a household will be more likely to select a supermarket within the commune that is the most visited by households within its commune

Parameter	2121	3232	4372	7694	NG11	NG17	R18	R31
disSM	-0.12	-0.15	-0.14	-0.20	-0.06	-0.05	-0.06	-0.04
dissqSM	.	.	.	0.00	0.00	0.00	0.00	0.00
surfSM	3.42	3.29	3.36	3.12	1.50	2.04	3.27	3.37
surfsqSM	-0.73	-0.68	-0.69	-0.61	.	.	-0.68	-0.69
samedepSM	0.72	0.78	0.76	0.79	2.46	2.30	1.41	1.38
sameuuSM	.	-0.16	.	-0.15	1.79	1.53	0.66	0.73
samecitSM	1.00	1.06	1.07	1.07	1.91	1.44	1.17	1.37
SMGStu299le0	-0.73	-0.92	-0.88	-0.97	-0.94	.	-0.86	-0.85
SMGStu299le1	-0.55	-0.75	-0.89	-0.90	.	.	-0.87	-0.64
SMGStu299le2	-1.45	-1.54	-0.78	-0.88	.	0.69	-1.81	-1.13
SMGStu299le3	-1.65	-2.00	-1.37	-1.60	.	-0.71	-2.31	-1.94
SMGSpol99le1	.	.	.	.	.	.	-0.80	.
SMGSpol99le2	.	.	.	.	.	.	.	0.22
SMGSpol99le3	-0.46	-0.55	-0.48	-0.42	.	.	-0.16	-0.52
SMGSVC99_1	-1.01	-1.15	-1.08	-1.11	-1.26	-0.72	-1.11	-1.13
TR2ROUSM	0.34	.	.	.	.	.	.	.
TRCOMSM	.	0.22	0.28	0.33	.	0.31	0.29	0.31
FavComSM	0.90	0.87	0.92	0.91	1.86	1.20	0.87	0.99
FavVilSM	.	.	0.80	0.71	.	.	.	0.68
AccComFSM	1.13	1.19	1.12	1.07	1.30	.	1.26	1.24
AccVilFSM	.	.	0.79	0.68	.	.	0.39	0.55
AccAutoRSM	-0.01	.	.	.	.	-0.01	.	-0.00
AccComFZSM	.	0.36	0.36	0.36	.	-0.96	0.18	.
AccVilFZSM	-0.13	.	0.41	0.36	.	0.40	.	.
AccAutoRZSM	-0.25	-0.26	-0.27	-0.17	-0.48	.	.	.
DensPopuSM	-0.27	-0.27	-0.32	-0.30	-0.13	-0.21	-0.37	-0.37

Table 3.7: Parameter Estimates Part 2: Comparison of parameters specific to supermarkets generated for various choice set definitions

of residence than a supermarket in another commune, all else being equal. That is, the more accessible the commune of the store is to the commune in which the household lives, the more likely it will be chosen. However, if "AccFavComSM" is positive it means that a supermarket will be more likely to be selected the further the commune is from the commune that is most visited by the households living in the same commune as the store. The second effect is not a contradiction of the first. Stores within a commune will be more likely to be chosen if they are more accessible to a household, and even more so if they are more isolated from their competition. Thus, "FavComSM" and "FavVilSM" add to notions of distance between domiciles and stores, and "AccFavComSM" and "AccFavVilSM" add to notions of the presence of competition with other large-surface stores. Thus, a commune's greater accessibility will be both a positive and a negative factor in the calculation of utility.

What most draws our attention here is the fact that the signs of the estimated coefficients of the variables "SameuuSM", "SameuuHM", "SameuuHD", and "SameuuXM" are negative when we use aggregated alternatives and positive when we use sampled alternatives. In both cases, these estimated parameters are significant, leading us to make different conclusions about shopping behaviour when we use different definitions of choice sets.

A positive value of "sameuuSM" may be unsurprising, since we would think that a household would be more likely to choose a store within the urban unit of its residence than one that is not within the same urban unit. However, we can also explain a negative value of "sameuuSM". Since a store within a household's commune is automatically in the household's urban unit, the value of "sameuuSM" is actually a measure of the difference in utility between a store that is within the same urban unit, and *not* within the same commune and a store that is outside the household's urban unit, all else being equal. If

this value is negative, then we can explain this by supposing that a household within a large city would find stores within the same part of the city to be in general more accessible (hence the reason why "sameCitSM" is positive", but would find a store in another part of the city less accessible than a store located just outside the city, since it would require the crossing of dense urban areas. Now, if this parameter is negative when we use aggregated choice sets, but positive when we use sampled alternatives to represent a consideration set encompassing all stores within the region, then we can conclude that this explanation only holds within the household's immediate neighbourhood. This difference in model parameters shows us that for at least when it comes to the effect of urban units on store choice, the proper definition of consideration sets matters.

Another delicate issue is the effect of urban density on a store's attractivity. The variables we choose permit us to separate contradictory effects associated with larger urban centres. On the one hand, "DensPopuSM" is negative, meaning that all else being equal, households prefer stores located within densely populated areas. However, "SMGStu299le0", "SMGStu299le1", and "SMGStu299le3" are all negative, meaning that households are more attracted to stores located in urban areas with larger populations. And they are also more attracted to stores that are in communes that are more economically central to their region, as attested by the negative value of "SMGSpol99le3". However, despite being more attracted to stores in larger urban areas, households are still less attracted to communes that represent the central part of an urban agglomeration, as shown by the negative value of "SMGSVC99_1". We can hypothesize that households will be more familiar with these stores that are near economic centres, since they are more likely to have passed near them on home-work, or socially-oriented trajectories. However, accessibility may be impeded by traffic congestion in more densely populated urban areas, and extra restrictions on travel may be imposed within down-

Parameter	2121	3232	4372	7694	NG11	NG17	R18	R31
disHM	-0.13	-0.16	-0.19	-0.26	-0.04	-0.03	-0.03	-0.03
dissqHM	.	.	0.00	0.00	0.00	0.00	0.00	0.00
surfHM	1.51	1.26	1.34	1.33	0.96	1.04	1.30	1.29
surfsqHM	-0.13	-0.12	-0.12	-0.12	-0.07	-0.09	-0.12	-0.12
samedepHM	0.94	0.99	0.91	1.01	2.25	2.41	1.56	1.43
sameuuHM	-0.62	-0.61	-0.59	-0.51	3.10	2.58	1.24	1.35
samecitHM	1.24	1.33	1.23	1.00	1.15	1.52	1.30	1.32
HMGStu299le0	.	.	.	.	.	.	0.77	0.89
HMGStu299le1	.	.	.	.	.	.	-0.55	-0.37
HMGStu299le2	-1.69	-1.66	-0.73	-0.65	.	.	-1.35	-1.01
HMGStu299le3	-1.02	-1.46	-0.65	-0.53	.	.	-1.38	-0.96
HMGSpol99le1	.	0.77	0.63	0.65	.	.	0.40	0.54
HMGSpol99le2	.	-0.98	-0.91	-0.89	.	.	.	.
HMGSpol99le3	-0.30	.	.	.	.	.	-0.90	-0.95
HMGSVC99_1	.	-0.22	.	-0.20	.	-0.43	-0.60	-0.50
TR2ROUHM	0.63	.	.	0.52	-0.79	.	0.66	0.65
TRCOMHM	-1.15	-0.46	-0.48	-0.61	.	.	-0.60	-0.53
FavComHM	1.16	0.96	0.91	1.11	1.88	1.83	1.55	1.63
FavVilHM	0.96	0.71	0.62	0.88	.	.	0.90	0.96
AccComFHM	.	0.35	.	.	.	.	0.73	0.44
AccVilFHM	1.60	.	.	-1.14	.	.	-1.54	-2.23
AccAutoRHM	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
AccComFZHM	-0.81	.	-0.30	-0.28	-1.33	-0.51	.	.
AccVilFZHM	.	-1.12	-1.05	-1.63	.	-0.69	-2.22	-2.52
AccAutoRZHM	0.29	0.22	.	0.18	.	.	.	.
DensPopuHM	-0.46	-0.23	-0.13	-0.21	.	.	-0.18	-0.30

Table 3.8: Parameter Estimates Part 3: Comparison of parameters specific to hypermarkets generated for various choice set definitions

town areas (pedestrian ways, one-way streets, etc) where streets may also be narrower. Because both economic centrality and population density are often associated with larger commune populations, it is important to be able to make use of this kind of detailed information to separate these effects. A reliance on only the population of a commune in order to predict the attractivity of a store within it may therefore give deceptive results. Such patterns of behaviour are generally confirmed in Tables 3.8 to 3.10.

In Table 3.11 we look at the effect of competition between supermarkets and hypermarkets that are located in the same commune. The variables in the table indicate

Parameter	2121	3232	4372	7694	NG11	NG17	R18	R31
disHD	-0.15	-0.17	-0.14	-0.15	-0.06	-0.04	-0.03	-0.02
dissqHD	0.00	0.00	0.00	0.00	.	.	.	.
surfHD	1.71	.	.	.	1.62	1.34	1.74	1.86
surfsqHD	.	0.86	0.84	0.75	.	.	.	.
samedepHD	0.64	0.61	0.82	0.72	2.22	2.27	1.42	1.48
sameuuHD	-0.81	-0.74	-0.48	-0.37	1.36	1.31	0.79	0.85
samecitHD	1.04	1.04	1.06	0.84	0.72	0.92	0.57	0.75
HDGStu299le0	-1.16	.	.	.	.	.	-1.43	-1.70
HDGStu299le1	0.67	-0.82	-0.72	-0.75	-0.68	.	.	0.79
HDGStu299le2	.	.	.	.	.	0.96	.	0.39
HDGStu299le3	.	.	0.38	0.21	.	-0.69	-0.40	-0.20
HDGSpol99le1	2.17	.	.	.	.	0.60	1.81	2.79
HDGSpol99le2	-2.34	.	-0.78	-0.75	.	.	-2.35	-2.57
HDGSpol99le3	.	-1.01	.	.	.	.	.	.
HDGSVC99_1	-0.58	-0.24	-0.35	-0.34	.	.	-0.61	-0.97
TR2ROUHD	.	-0.47	-0.32	.	.	.	.	.
TRCOMHD	-0.16	.	.	.	.	.	.	.
FavComHD	0.75	0.71	0.74	0.74	0.63	.	0.71	0.82
FavVilHD	.	.	0.46	0.51	.	.	.	0.43
AccComFHD	1.03	0.94	1.11	0.94	.	1.00	1.33	1.35
AccVilFHD	-3.21	-4.00	-2.69	-3.09	.	.	-4.13	-2.85
AccAutoRHD	.	.	.	.	-0.03	-0.05	-0.01	-0.01
AccComFZHD	.	.	0.34	0.38	.	.	.	.
AccVilFZHD	-2.71	-3.08	-2.22	-2.48	.	.	-3.31	-2.50
AccAutoRZHD	.	-0.32	-0.50	-0.58	-0.99	-1.11	-0.28	.

Table 3.9: Parameter Estimates Part 4: Comparison of parameters specific to hard discounts generated for various choice set definitions

Parameter	2121	3232	4372	7694	NG11	NG17	R18	R31
disXM	-0.16	-0.16	-0.17	-0.18	-0.09	-0.09	-0.12	-0.11
dissqXM	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
surfXM	0.43	0.05	0.34	0.30	.	.	2.08	1.65
surfsqXM	-0.02	.	-0.01	-0.01	-0.02	.	-0.09	-0.07
samedepXM	.	-0.56	.	-0.72	.	.	-0.50	-0.49
sameuuXM	-0.92	-0.70	-0.66	-0.43	1.89	1.94	0.34	0.37
samecitXM	1.12	1.15	1.15	0.99	1.32	1.44	0.97	1.08
XMGStu299le2	-1.98	-2.58	-1.95	-2.59	.	.	-2.47	-2.46
XMGStu299le3	.	0.39	0.38	0.47	.	.	0.32	0.35
XMGSpol99le1	.	.	.	.	1.94	1.18	.	.
XMGSpol99le2	.	.	.	.	.	.	-15.07	-11.32
XMGSVC99_1	.	0.49	.	0.28	.	0.68	-0.34	-0.37
TR2ROUXM	-2.39	-2.85	-2.96	-2.70	.	.	-2.23	-2.25
TRCOMXM	1.05	1.25	1.36	1.37	.	.	1.26	1.24
FavComXM	.	.	.	.	.	0.78	0.41	0.40
FavVilXM	0.65	0.62	0.54	0.59	.	.	0.89	0.89
AccComFXM	3.22	.	1.69	.	.	.	.	.
AccAutoRXM	.	.	.	.	-0.06	.	.	.
AccComFZXM	.	.	.	.	1.33	.	.	.
AccAutoRZXM	0.87	.	0.65	.	.	.	0.29	0.33
DensPopuXM	-0.29	-0.13	-0.24	-0.14	.	.	-0.22	-0.21

Table 3.10: Parameter Estimates Part 5: Comparison of parameters specific to large hypermarkets generated for various choice set definitions

the number of hypermarkets and supermarkets within the same commune. These variables serve as simple proxies for spatial correlation. If there is spatial autocorrelation between stores, store utilities will be correlated with the utilities of nearby stores, all else considered equal. A conditional Logit model can be modified to take into account such effects, notably by introducing correlation coefficients in the distribution of the error terms of the model. This is a theoretically sound practice, and has been explored in the developement of generalizations of Conditional Logit models such as in Generalized Extreme Value (GEV) models, for example (Train, 2003, pp 80-100 [46], Guo, 2004 [16]). We do not enter these sorts of effects in our model, due to the increased computational burden, but we can take into account the effect of competition between nearby stores. The variable "SM_CntSMGE3" indicates that a given supermarket is located in a commune with at least 3 other supermarkets. We see that the estimated coefficients of this variable are negative, meaning that a household will be less likely to choose a supermarket if it is located in a commune that contains at least two other supermarkets, all else being equal. The variable "SM_CntSM1HM0" is one when the alternative is the choice of a supermarket in a commune containing no small hypermarkets and no other supermarkets.

A look at this table will lead one to conclude that the probability of selecting a given supermarket and hypermarket declines the more supermarkets and hypermarkets are in the same commune and therefore in competition with it. Interestingly, the probability of selecting a supermarket or hypermarket actually increases if there are hard discount stores or large hypermarkets inside the commune. We note that the use of spatial correlations could permit a more sophisticated understanding of competition between stores, since it will be take into account not only the number of stores within the same commune, but the utilities of these stores and their distances from the store in question.

Parameter	2121	3232	4372	7694	NG11	NG17	R18	R31
SM_CntSMGE3	-0.44	-0.30	-0.24	-0.17	.	.	.	-0.27
SM_CntHMGE2	-0.92	-0.66	-0.77	-0.87	.	.	-0.69	-0.80
SM_CntHMGE3	-1.28	-2.08	-1.41	-1.44	.	.	-1.47	-1.17
SM_CntXMGE1	-1.52	-2.17	-1.58	-1.64	.	.	-1.95	-1.53
SM_CntSM1HM0	0.41	0.33	0.31	0.31	0.94	.	0.39	0.32
SM_CntSM2HM1	-0.97	-0.87	-0.88	-0.89	.	.	-1.15	-1.25
HM_CntSMGE2	-0.40	-0.45	-0.38	-0.37	.	.	-0.47	-0.52
HM_CntSMGE3	0.70	0.71	0.83	0.81	.	.	0.69	0.80
HM_CntSMGE4	-1.29	-1.79	-1.37	-1.36	.	-0.54	-1.34	-1.57
HM_CntHMGE2	-0.65	-0.39	-0.27	-0.20	-0.99	.	-0.54	-0.65
HM_CntHMGE3	-0.80	-1.74	-1.13	-1.02	.	.	-0.72	-0.43
HM_CntHDGE2	0.84	1.00	1.10	0.95	.	.	.	0.37
HM_CntXMGE1	.	-0.54	.	0.38	.	.	-0.41	.

Table 3.11: Parameter Estimates Part 6: Effects of competition

3.1.3 Evaluating the Logit Models

Once we have developed our Logit model, and estimated its parameters, we need to be able to evaluate our model. We need to be able to determine which of the many choices we had made in our model specification produced the best results, and be able to decide how effective our model is for its cost. We look at three ways in which our model is evaluated through the McFadden Pseudo R-squared term, the Brier Score, and a score we developed that we call the "WD" statistic.

McFadden R-squared calculation

We introduced the McFadden R-squared in Section 2.2. Here, we calculate the McFadden R-squared term on each of the four choice set definitions for the aggregated alternatives model. We would like to see how this measure works in the comparison of a model of choice behaviour applied to the same data, but with slightly different choice set definitions. In our data on choices of large-surface stores, we define the following four choice sets listed in Table 3.12:

2121	3232	4372	7694
SM1	SM1	SM1	SM1
SM2	SM2	SM2	SM2
	SM3	SM3	SM3
		SM4	SM4
			SM5
			SM6
			SM7
HM1	HM1	HM1	HM1
	HM2	HM2	HM2
		HM3	HM3
			HM4
			HM5
			HM6
HD1	HD1	HD1	HD1
HD2	HD2	HD2	HD2
	HD3	HD3	HD3
		HD4	HD4
		HD5	HD5
		HD6	HD6
		HD7	HD7
			HD8
			HD9
XM1	XM1	XM1	XM1
	XM2	XM2	XM2
			XM3
			XM4
OUT1	OUT2	OUT3	OUT4
NO	NO	NO	NO

Table 3.12: Choice Sets

In each choice set, each of the first alternatives refers to different large-surface store choices except the last two. "NO" refers to the choice of no large-surface store, and "OUT" refers to the outside option, that is, a choice of large-surface store not already enumerated in the choice set. The outside option is different in each choice set. If an individual chooses "XM2", this would be considered a selection of "Out1" if we use "2121", but not if we use "3232", "4372", or "7694". We compare the McFadden R^2 calculated on a Conditional Logit Model run using all four of these definitions:

Value	2121	3232	4372	7694
LL_1	-16700.49	-20190.72	-21919.08	-23800.94
LL_0	-21022.11	-25705.12	-27818.96	-30114.34
K	88	88	95	104
K'	8	12	18	28
R2	0.2056	0.2145	0.2121	0.2096

Table 3.13: Log Likelihood and R^2 by choice set

Although the sum of the log likelihoods for each alternative are quite different, the McFadden R^2 terms listed in Table 3.13 are close to identical for each choice set definition. We recall that there is a set of alternatives C that applies to every single individual in a sample, the formula for a log-likelihood expression is

$$\sum_{n=1}^{N}\sum_{j\in C} z_{nj}\ln P_{nj} \;=\; \sum_{n=1}^{N} z_{n1}\ln P_{n1} + \sum_{n=1}^{N} z_{n2}\ln P_{n2} + \cdots + \sum_{n=1}^{N} z_{nj}\ln P_{nj} + \cdots$$

This means that we can break our log likelihood expressions into sums of likelihoods by alternative, which we do in Table 3.14, where we look at the estimated log likelihood of each alternative selected separately:

Here, the values in line "Out1" represent the estimated log likelihood of households selecting the alternative "Out1". "Out1D" is the difference between the log likelihood

Value	Choice	2121	3232	4372	7694
LL_1	Out1D	0	-3511	-5230	-7131
	Out1	-4102	-4035	-4033	-4041
	SM1	-3239	-3220	-3224	-3223
	SM2	-1908	-1907	-1899	-1895
	HM1	-2861	-2906	-2911	-2894
	HD1	-1015	-1024	-1031	-1027
	HD2	-609	-625	-621	-623
	XM1	-2098	-2096	-2100	-2096
	NO	-867	-867	-869	-870
LL_0	Out1D	0	-4693	-6806	-9102
	Out1	-4586	-4579	-4580	-4580
	SM1	-4313	-4309	-4305	-4304
	SM2	-2442	-2437	-2436	-2436
	HM1	-3951	-3954	-3956	-3954
	HD1	-1037	-1037	-1037	-1037
	HD2	-654	-654	-654	-654
	XM1	-3155	-3157	-3160	-3163
	NO	-884	-884	-884	-884

Table 3.14: Log Likelihood by alternative and choice set

of the alternatives included within the definition of "Out1" and the log likelihood of selecting the alternative "Out1". In effect this is the difference between the log likelihood of a given model if we aggregate the predicted probabilities that are in "Out1" and the log likelihood if we don't. For example, we can find the log likelihood of LL_1 for the model "3232" by summing the values in the first part of the table, under 2121. This gives us a value of -20190.72. We note that the alternatives "SM3", "HM2", "HD3", "XM2", and "Out2" are considered separate alternatives in "3232", but are included in the alternative "Out1" in "2121". We could create a value of the log likelihood that would be more comparable to the one assigned to "2121" by assigning the sum of the probabilities of selecting all alternatives classified as belonging to "Out1" to the household's choice of store whenever it selects one of these alternatives. The sum of the logs of these new assigned probabilities for the model "3232" is found in the second term of the second row of our table, -4035. This table shows us that the log likelihoods of the models defined for

Choice	2121	3232	4372	7694
SAS R2	0.3409	0.3355	0.3806	0.4158

Table 3.15: McFadden Pseudo-R-squared term calculated by SAS.

the four choice sets for both the null and fitted models differ only in the log likelihoods assigned to alternatives that are included in "Out1".

We would like to add a warning to SAS users who have used the MDC Procedure. In Table 3.15, we show the values of the McFadden 4-squared terms that are calculated by this procedure. Unfortunately, what could lead to confusion is the fact that the MDC R^2 is not the one we define, nor does it correspond to the one mentioned in Domencich and McFadden[11]. The SAS index takes as the null model, a model that assumes equal utilities across all individuals and all alternatives, which is equivalent to an average log likelihood of the function.

$$\text{SAS R}^2 \;=\; 1 - \frac{\text{LL}}{\sum_{n=1}^{N} \frac{1}{|C_n|}}$$

where $|C_n|$ is the number of terms in the set C_n. If this is equal to J for each individual, then

$$\text{SAS R}^2 \;=\; 1 - J \times \frac{\text{LL}}{N}$$

The fact that the R^2 terms rise as we increase the number of alternatives is simply a reflection of the more unbalanced nature of the probabilities assigned to our alternatives as we increase sample size. If the null likelihood in SAS is directly proportional to the inverse of the log of the number of alternatives in each individual's choice set, it can be seen from Table 3.14 that this is not the case for our null likelihoods, which change as different proportions of small probabilities are included in the model's choice set.

We did not calculate a McFadden Pseudo-R-squared term for the models in which we sampled alternatives. For these models, we would need a meaningful definition of a null model. Unfortunately, if we take as our choice set, the set of all stores within the region, many of our null probabilities, that is, the proportion of households in our population selecting each store would be zero, rendering a null log likelihood term indefinite.

Predictive Accuracy

Using these estimates of model parameters, we can forecast the probabilities of selection of every alternative associated with each individual and each choice set. In Table 3.16, we compare the "WD" statistics (described in Section 2.2) that we calculate over each model using different geographic subdivisions of the Centre Region. Forecasting the percentage of households visiting each store when using aggregated probabilities is complicated by the fact that a prediction of the selection of the "outside" option cannot be attributed to any single store. We also see that the "WD" statistic is smaller, the more narrowly we define the geographic zone upon which it is calculated. As a comparison with our "WD" statistic, we also calculate the "Calibration" of our model (described in Section 2.2.2) over the survey sectors.

Statistic	Zone	2121	3232	4372	7694	NG11	NG17	R18	R31
WD	Centre	0.710	0.797	0.825	0.851	0.635	0.592	0.625	0.657
WD	Secteur	0.648	0.714	0.733	0.749	0.534	0.505	0.555	0.578
WD	IRIS	0.585	0.644	0.659	0.670	0.464	0.444	0.515	0.523
Calib	Secteur	0.403	0.187	0.118	0.083	4.495	5.616	2.829	2.971

Table 3.16: Predictive accuracy of different Logit models.

From this table, we see that using aggregated probabilities is clearly superior to using sampled alternatives. There are a few possible explanations for this difference. The first is simply the inadequacies of our estimation method. Although it has been established

that our estimations are asymptotically consistent with those calculated using the full choice set, our estimations using sampled alternatives can still be far from those using the full set of choice alternatives. The second is that the assumption that all large-surface stores are part of each households' consideration set may be unreasonable. Having such a huge number of large-surface stores with such an enormous variety of distances from the households stretches the plausability of the independence of irrelevant alternatives (IIA) hypothesis. It may be reasonable to believe, for instance, that introducing a new large-surface stores in a choice between several nearby stores would decrease the probability of selecting each other store proportionally. Unfortunately, this proportionality may be harder to accept when we are comparing stores that are near with those that are so far away that their attraction on the household takes on a categorically different quality.

We also see that in the aggregated choice models "2121", "3232", "4372", and "7694, that the accuracy of predictions increases as the choice set defined for each household increases in size. This is natural, since there is no one store in our sample that corresponds to the "outside" option. This means that the greater the expected number of households choosing the outside option, the less households will be left to be distributed amongst actual store choices. Since the "2121" assigned much greater probabilities to the outside option than the other models, the number of households it predicts would visit each store is further from the actual number than in the other models. However, we must keep in mind that the accurate prediction of all flows between all households and all stores may not interest the statistician. Perhaps we are only interested in the most common store choices, and the number of households who leave their immediate neighbourhood to go shopping. If we recalculate the "WD" statistic, calling it "WDNO", in which we treat all selections of the outside option as if the outside option were another store in the region, we obtain the results for the different aggregated models in Table 3.17. In this case,

because the "2121" model predicts more households in the outside option, the values of "WDNO" are now better for "2121" than for the other models.

Statistic	Zone	2121	3232	4372	7694
WDNO	Centre	0.935	0.920	0.909	0.900
WDNO	Secteur	0.895	0.849	0.823	0.801
WDNO	IRIS	0.849	0.787	0.755	0.726
CalibNO	Secteur	0.094	0.093	0.106	0.102

Table 3.17: Predictive accuracy of different models of aggregated choices, outside option considered as a store in its own right.

	2121	3232	4372	7694	NG11	NG17	R18	R31
Estimation	01:00	02:35	08:28	13:29	03:01	07:33	01:45	03:50
NObs	102473	153107	229202	354621	150365	225121	150156	221133
MB	24.7	39.2	61.2	96.9	23.1	35.2	32.2	47.0
Params	88	88	95	104	52	62	98	105
Forecast	00:01	00:01	00:00	00:01	00:07	00:04	00:06	00:05
NObs	102473	153107	229202	354621	2601648	2601648	2601648	2601648
MB	3.8	5.6	8.4	12.5	109.6	109.6	109.6	109.6
ExpCen	00:00	00:00	00:00	00:00	00:06	00:05	00:06	00:07
ExpSec	00:00	00:00	00:00	00:00	00:05	00:03	00:05	00:04
ExpIri	00:00	00:00	00:00	00:00	00:17	00:17	00:13	00:14

Table 3.18: Comparison of Computational Burden of different Logit Models.

We also compare the different models in terms of the computational resources required. In the first row of Table 3.18, we compare the time in hours and minutes required in order to obtain the estimates of all the parameters of our model. In the following two rows, we compare the SAS files we use for the model parameter estimates in terms of the number of observations they contain, and the number of megabytes of space they occupy. We also compare the number of significant parameters amongst the model's parameter estimates. We can see that the estimation time rises rapidly as we increase the number of alternatives that we include in each choice set. For forecast probabilities, once again, we calculate the calculation time required to obtain the files containing

the forecast probabilities of selection of each alternative in each choice set, along with the number of observations and the file size in megabytes of the files used in order to calculate forecast probabilities. In the last three lines of Table 3.18, we also provide the calculation times to obtain the forecast number of households from each geographic zones choosing each large-surface store. We do this for the number of households coming from the entire region, from each survey sector, and for each IRIS, respectively. When we look at the model forecasts, we see that once the parameter estimates are determined, forecasting model probabilities using these estimates goes fairly quickly, even for large data sets. However, we see that using sampled probabilities, we must predict probabilities of selection for all large-surface stores in the region for each household. Such a large number of predicted probabilities (most of whose values are extremely small) can become cumbersome, as seen by the recorded times required to use these probabilities to forecast the number of households from each geographic zone visiting each large-surface store. This is a serious issue to be kept in mind when comparing different models, for it will add to the cost of a model's use.

In order to do cross-validation by department, we calculate estimated coefficients using five different subsets of the original data set. In Table 3.19, we compare the parameter estimates for the "2121" model obtained by using our estimation over all households in the entire sample and compare these with the estimates obtained with five training samples each obtained by taking households in all but one of the departments. As we can see, the parameter estimates differ greatly in the number of parameters that are treated as significant in the model. Many parameters that are significant in one estimation are not in another. However, where a parameter is significant in more than one model, it generally has the same sign.

Rather than printing all parameter estimates for all training sets for all models of

Parameter	Original	Not 18	Not 28	Not 36	Not 37	Not 41
SM	1.476	.	-2.407	4.507	-2.060	-1.497
SMRankGE2	-0.339	-0.364	-0.326	-0.355	-0.230	-0.385
HDRankGE2	-0.223	-0.281	.	-0.468	.	-0.248
XM	.	.	-1.612	.	.	.
outside	1.161	.	.	2.957	-0.414	.
nostore	-1.766	-2.891	-2.952	.	-3.348	-2.898
SM_CntSMGE3	-0.438	-0.406	-0.341	-0.604	-0.371	-0.400
SM_CntHMGE2	-0.925	-0.873	-1.099	-2.224	-1.313	-0.414
SM_CntHMGE3	-1.279	-1.500	0.379	.	-1.044	-1.886
SM_CntXMGE1	-1.518	-1.536	.	-1.688	-1.457	.
SM_CntSM1HM0	0.413	0.469	0.445	0.378	0.589	0.273
SM_CntSM2HM1	-0.966	-0.904	-0.811	-2.236	-0.962	-0.944
HM_CntSMGE2	-0.401	-0.476	-0.474	.	0.363	0.662
HM_CntSMGE3	0.699	1.276	0.717	.	-0.590	0.983
HM_CntSMGE4	-1.288	.	-2.009	-1.901	.	.
HM_CntHMGE2	-0.652	-1.205	-0.826	-0.996	.	-0.368
HM_CntHMGE3	-0.795	.	0.422	-1.487	-2.022	.
HM_CntHDGE2	0.836	0.808	0.938	1.804	0.846	1.338
HM_CntXMGE1	.	1.167	1.540	-2.218	-1.085	.
disSM	-0.117	-0.119	-0.119	-0.125	-0.131	-0.121
surfSM	3.422	3.315	3.708	3.351	3.220	3.276
surfsqSM	-0.730	-0.700	-0.821	-0.714	-0.691	-0.669
samedepSM	0.719	0.805	0.701	0.802	0.764	0.727
sameuuSM	.	.	.	.	-0.196	.
samecitSM	1.000	1.076	1.039	0.991	0.988	0.875

Table 3.19: Parameter estimates of "2121" model contrasted with training set estimates for cross-validation.

aggregated choices, we create Table 3.20 that shows the stability of these estimates calculated on different data sets. Under "Orig Parm" we print the number of parameters retained as significant in the parameter estimates done over the entire survey region. Under "Total Parm" we count the total number of parameters retained as significant in an estimation done over at least one training set. Under "New Sig" we count the number of times that a parameter not considered significant in an estimation done on one of the training sets is in fact significant in an estimation done on an estimation done over the entire survey region. Under "New Sgn" we count the number of times that a parameter estimate has a different sign on the training set than it does on the original estimate over the entire survey region. Under "Low Var" we calculate whether there is a great deal of difference between the parameters calculated on the training sets and the original parameter estimates. We do this by calculating the square root of the average squared difference between the training set parameters and the original parameter estimate, and divide this by the absolute value of the original parameter estimate. Under "Low Var" therefore, we count the number of parameters for which this value is less than one, and thus what we consider relatively stable.

Once we have obtained predicted probabilities for all households in the region, we can calculate the "WD" statistics exactly as before. This is represented in Table 3.21. The difference between the values obtained in this table can be attributed to the elimination of overfit.

We also calculate the value of the "WDNO" statistics after cross-validation, in Table 3.22.

We can now see the importance of using cross-validation so as not to get an over-estimate of the performance of our model when applied to new data sets. In both the ordinary "WD" calculation and in the "WDNO" where we are only validating predictions

Order	Model	Orig Parm	Total Parm	New Sig	New Sgn	Low Var
First	2121	82	111	326	24	63
	3232	85	113	348	13	72
	4372	92	121	382	17	70
	7694	108	133	441	25	88
Second	2121	80	109	313	13	59
	3232	84	111	339	20	67
	4372	88	118	355	17	71
	7694	101	131	413	25	83
Third	2121	54	95	186	16	40
	3232	67	99	262	16	54
	4372	74	109	275	15	56
	7694	82	110	326	16	62

Table 3.20: Stability of model parameters in cross-validation

Statistic	Zone	2121	3232	4372	7694
WD	Centre	0.628	0.690	0.669	0.714
WD	Secteur	0.556	0.628	0.612	0.645
WD	IRIS	0.501	0.573	0.556	0.584
Calib	Secteur	0.596	0.270	0.266	0.203

Table 3.21: Cross-validation of model results for aggregated choice sets.

Statistic	Zone	2121	3232	4372	7694
WDNO	Centre	0.833	0.833	0.789	0.809
WDNO	Secteur	0.750	0.753	0.694	0.708
WDNO	IRIS	0.713	0.702	0.640	0.645
CalibNO	Secteur	0.456	0.196	0.245	0.197

Table 3.22: “WDNO” statistics after cross-validation.

of visits of close large-surface stores, there is a great difference between the evaluative criteria with and without cross-validation. We can see here that our model is sensitive to the geographic region to which it is adapted. We can see that we can obtain relatively good forecasts of store market shares even while drastically reducing the cost of estimation. Our cross-validation does show, however, that as in the case of the "4372" a larger choice set may not necessarily lead to improved "WD" statistics.

3.2 Gravitational Model Estimation

The Conditional Logit models described above, due to the enormous number of explanatory variables included, are very cumbersome to estimate. We therefore contrast this model with the very simplest model that we can develop, which we call the "gravitational model". According to this model, every single choice of store j for each individual n will be assigned a weight that will depend on the store type, (supermarket, hypermarket, hard discount, and large hypermarket), distance d_{nj} and the store's "mass" which we shall take to be its retail space s_j). The weight will be calculated according to the following formula:

$$A_{nj} = \frac{s_j^{\alpha(t(j))}}{d_{nj}^{\beta(t(j))}}$$

where $t(j)$ refers to the type of large-surface store j where it is "SM", "HM", "HD", or "XM", leaving us with 8 parameters, α_{SM}, α_{HM}, α_{HD}, α_{XM}, β_{SM}, β_{HM}, β_{HD}, and β_{XM} that we need to estimate.

These weights can be used to estimate the probability of selection of an alternative

through the formula:

$$P_{nj} = \frac{A_{nj}}{\sum_{k \in U_n} A_{nk}}$$

As we explain in Appendix 5, we find that a regression done without using the retail space in our calculation fits better with the data than includes retail space. We note that because of the vast number of stores included in our sample including just about all stores less than 100 km of any household in the survey area, the geometric mean of all distances in each choice set remain close to constant. We see this by looking at the quantiles of the average of the logs of the distances for the set of stores corresponding to each IRIS in Table 3.23.

Quantile	SM	HM	HD	XM
100%	12.04	12.11	12.10	12.23
99%	12.01	12.08	12.07	12.18
95%	11.97	12.04	12.01	12.13
90%	11.94	12.00	11.98	12.08
75%	11.84	11.91	11.89	11.94
50%	11.72	11.77	11.75	11.73
25%	11.66	11.73	11.66	11.59
10%	11.64	11.72	11.56	11.40
5%	11.64	11.70	11.54	11.32
1%	11.63	11.69	11.52	11.28
0%	11.63	11.67	11.50	11.23

Table 3.23: Quantiles of average of logs of distances between households and stores.

Once we have assumed that the denominator of the expression of our selection prob-

ability is constant in our sample, we simply write

$$\begin{aligned} P_{nj} &= \frac{d_{nj}^{\beta}}{\sum_{k \in C_n} d_{nk}^{\beta}} \\ &= \frac{d_{nj}^{\beta}}{K} \end{aligned}$$

which, when taking the log of both sides, yields

$$\ln P_{nj} = \alpha + \beta \ln d_{nj}$$

In order to estimate this expression, we need to approximate P_{nj}. We do this by taking the cartesian product of all households and all large-surface stores, and dividing this into a set of groups g of household-store-choice pairs such that if $(n, j) \in g$,, $d_{nj} \approx d_g$. Thus, for all n and all j, if $(n, j) \in g$, then

$$\begin{aligned} \ln P_{nj} &\approx P_g \\ &= \alpha + \beta \ln d_g \end{aligned}$$

Noting that

$$\sum_{g \in G} \frac{1}{\text{var}(P_g)} (\ln P_g - \alpha - \beta \ln d_g)^2$$

is a sum of squares of centred and homoskedastic random variables, we can estimate our parameters by finding the values of α and β that maximize this expression. This is a weighted least squares estimate. Since P_g is unknown, we estimate it with O_g, where

$$O_g = \frac{\sum_{(n,k) \in g} z_{nk}}{|g|} \tag{3.1}$$

Once again, z_{nk} is one when n selects k and zero, otherwise. Thus, our parameter estimates are the approximate weighted least squareds estimate maximizing the following function:

$$\sum_{g \in G} \frac{1}{\text{var}(O_g)} (\ln O_g - \alpha - \beta \ln d_g)^2$$

Here, we need to define the classes g in such a way that O_g is positive for all values of g. If the number of members of each class g is large enough, we can replace $\text{var}(O_g)$ with $|g|$, the number of household-store pairs in class g. Otherwise, we can estimate the variance with the formula:

$$\text{var}(O_g) = |g| O_g (1 - O_g)$$

For our calculations, household-store choice pairs are divided into classes by dividing the log of the recorded distances into regular intervals having on average K store selections recorded for each interval. We need to keep the intervals of the logs of the distances regular in order to ensure that the variance of the logs of the distances contained within each class remains constant. We also needed to keep these intervals large enough to exclude the possibility of having classes that contain no observations, since the fact that they were excluded from our regression could bias our results, eliminating data points corresponding to very small values of P_g.

The problem with this expression, unfortunately, is that it does not define a probability, since extreme values of d_{nj} could lead to negative values of P_{nj}. Fortunately, for almost all observations in our data set, we do not have negative values of P_{nj}.

The results of our parameter estimates through weighted least squares regression, along with the R-squared values of each model are in Table 3.24 (with representing the average number of households per interval of the log of the distance):

H	store type	R^2	Intercept	Avg ln(dis)
50	SM	0.616	8.471	-1.621
	HM	0.788	16.964	-2.203
	HD	0.769	13.534	-2.147
	XM	0.892	14.933	-1.994
20	SM	0.572	10.030	-1.822
	HM	0.760	20.572	-2.572
	HD	0.779	20.751	-2.870
	XM	0.870	20.284	-2.557
5	SM	0.555	11.397	-1.950
	HM	0.715	22.173	-2.728
	HD	0.852	24.934	-3.243
	XM	0.784	19.921	-2.565

Table 3.24: Regression estimates of gravitational model parameters

We see here that the more data points we use, (and thus, the smaller the number of households upon which the averages of each interval are calculated), in general, the lower the R^2 term are (as usual, the hard discount stores are an exception). However, a better model fit over averages of larger groups may not reflect parameters that are better fit to the individuals in the population, and this means that the only way we can be sure to determine which model parameters work best is to use them to calculate predicted probabilities of selection and then evaluate these according to the uses we have set for them. We do note that when we use only household-store-choice pairs corresponding to far stores, the parameters change. This should come as no surprise since we have already seen in our Conditional Logit estimation the large number of variables upon household decision-making depends. If we wish to improve our model, we could fit more complex curves to our data points. However, there would be little point in following up on this if we have already developed a Conditional Logit Model which we believe to provide the best means of generating probabilities of selection.

The gravitational model we have developed is conditionned on store type. Therefore,

if we want to use it to assign probabilities to choices of large-surface stores, we will need to assign a probability of selecting each store type. The probability of n selecting the large-surface store j will thus be written

$$P_{nj} = \left\{ P^T(t(j)) \frac{A_j}{\sum_{k \in C_n} A_k \delta_{t(k)=t(j)}} \right. , j \in C_n$$

where $t(j)$ refers to the store type of store j and $\delta_{t(k)=t(j)}$, is an indicator function that is one when k and j are of the same store type and zero, otherwise.

We originally intended to assign the probability P^T of selecting each store type using a multinomial logit model. The explanatory variables we would use would be the sociodemographic data that we had at our disposal describing each household in detail. All these variables would be crossed with the four alternatives specified for each individual: choice of supermarket, choice of hypermarket, choice of hard discount, and choice of large hypermarket. We attempted a model in which we would progress iteratively, eliminating nonsignificant variables one by one. We abandoned this technique, since we had found that it did not improve the results of our estimation enough to make it worthwhile. This once again confirmed our earlier assertions that sociodemographic variables on their own were not adequate predictors of store choice. Such efforts were simply not worthwhile when we were looking at marginal improvements in choices that already had very low probabilities. We therefore simply assigned to each household the same probabilities of selecting each store type in the outside option which we obtained by calculating the market shares of each store type.

We have tested some of the parameters generated through our regression calculations above by using them in order to define probabilities of selection of each choice of large-surface store for each household, and then used these to calculate a well-distributed statistic by sector over all three choices of large-surface stores. We have these in Table

3.25.

Stat	WD	WDIn
ContLin	0.365	0.086
N05Lin	0.581	0.408
N20Lin	0.610	0.451
N50Lin	0.621	0.467
N100Lin	0.618	0.463

Table 3.25: Well-distributed statistic for gravitational model, by different choices of gravitational parameters

The first line of our table is a control model, in which we suppose that the probabilities of any individual selecting any store is equal to the proportion of households selecting the store in the entire sample. This is a sort of "null" model serving as a baseline case. Because we have included second and third choices of large-surface stores in order to calculate the well-allocated statistic, a large number of household store choices are actually nonchoices. In order to see how many accurate predictions of actual individual stores are made, we calculate the "WDIn" statistic in which the predicted and observed nonchoice options were set to zero.

We begin by comparing all the parameters of our gravitational model generated for K equal to 5, 10 , 20, and 50 and find that for the purposes of our model, the parameters of our gravitational model generated using a linear regression on empirical probabilities of selection associated with household-store choice pairs having on average 50 individuals will work the best.

3.3 The "Hybrid" modeling strategy

We have seen in earlier sections that the aggregated choices model worked better than the sampled alternatives model in terms of the calculated "WD" statistic. However,

we also remarked that these types of aggregated models have the disadvantage of not accounting for low-probability choices of large-surface stores that are very distant. The assignment of an outside option is in fact not very meaningful. If we wish to predict the number of clients visiting each store in the sample, all households assigned an outside option will be assigned to no store. This means that the smaller the stores we include in the outside option, the more we will tend to underestimate the number of clients visiting each store. In order to rectify this, we use a simplified model in order to assign store choices to households assigned an outside store choice.

In order to assign store choices to households, we use the more sophisticated, yet hopefully more accurate conditional logit model to predict the probability that a household will select the few stores that are closest to its home, and then use the much more simple gravitational model in order to assign the probabilities that the household will select any other store conditional on the household choosing the "outside" alternative. Due to the fact that the vast majority of households select stores that fall into the set of stores whose probability of selection is modeled by the conditional model, the greater imprecision of the gravitational model should have minimal effect on the assigned store choices.

j in the universal set of elemental alternatives U_n:

$$P_{nj} = \begin{cases} P_j^{CL} & , j \notin o_n \\ P_{o_n}^{CL} P^T(t(j)|o_n) \dfrac{A_j}{\sum_{k \in o_n} A_k \delta_{t(k)=t(j)}} & , j \in o_n \end{cases}$$

Here, P^{CL} refers to the probability of the aggregated choice model, and o_n is the alternative in the aggregated choice model corresponding to the "outside" option. The function t indicates the type of store of j. As before, $\delta_{t(k)=t(j)}$ is an indicator function that is one when k is a store of the same type as j and zero otherwise. $P^T(t(j)|o_n)$ is the

Store Type	7694	4372	3232	2121	All
SM	46.5%	47.8%	47.4%	40.1%	45.8%
HM	41.7%	34.3%	33.9%	38.0%	30.6%
HD	6.9%	5.6%	9.7%	7.9%	5.8%
XM	4.9%	12.4%	9.0%	14.0%	16.0%
Total	1037	1641	2262	3954	12573

Table 3.26: Breakdown of types of store choices within choices of the outside option for different choice set definitions.

probability that household n selects a large-surface store of the same type as j conditional on the household's choice of store being within the outside option. The probability of a household selecting a store conditional on the store type and conditional on the choice being within the outside option will now simply be the fraction of weights in the outside option and having the same store type represented by this store. The weights A_j are of course the weights assigned by the "gravitational" model in the previous section.

We note that the breakdown of the households selecting a store in the outside option by type of store they selected will vary depending on how the outside option is defined. Table 3.26 shows the proportion of households selecting stores of each type when they select a store in the outside option.

The values in Table 3.26 correspond to our estimates of the values of P^T. When we wish to do cross-validation, we shall use the households in the training set choosing stores within the outside option in order to determine the values of P^T that we will apply to the test set.

In Table 3.27, we compare the "WD" calculated for the outside options of each model using the gravitational model. We compare this model with the results brought about by the gravitational model on its own and a control case where we assign the closest store to each household for the first choice, and no store for their second and third choices.

Hybrid Model				
	2121	3232	4372	7694
Centre	0.842	0.859	0.866	0.873
Sector	0.714	0.738	0.745	0.751
IRIS	0.635	0.659	0.666	0.671
Aggregated Alternatives Model				
	2121	3232	4372	7694
Centre	0.710	0.797	0.825	0.851
Sector	0.648	0.714	0.733	0.749
IRIS	0.585	0.644	0.659	0.670
Sampled Alternatives Model				
	NG11	NG17	R18	R31
Centre	0.635	0.592	0.625	0.657
Sector	0.534	0.505	0.555	0.578
IRIS	0.464	0.444	0.515	0.523
Gravitational Model				
Centre	0.763			
Sector	0.614			
IRIS	0.547			
Control Case				
Centre	0.541			
Sector	0.498			
IRIS	0.465			

Table 3.27: Comparison of the results of different store selection probability assignment strategies.

We see, as expected, that the "WD" statistics for each of the "Hybrid" model is somewhere in between the "WD" and the "WDNO" statistics in the aggregated choice models.

And once again, in Table 3.28, we redo the calculations, this time using cross-validation by department. We see that in cross-validation, the accuracy of model predictions drops substantially. This, we believe can be attributed to an inability to predict the type of store chosen in the outside option of each household in question.

Hybrid Model				
	2121	3232	4372	7694
Centre	0.692	0.726	0.695	0.726
Sector	0.593	0.643	0.618	0.646
IRIS	0.529	0.582	0.559	0.584
Aggregated Alternatives Model				
	2121	3232	4372	7694
Centre	0.628	0.690	0.669	0.714
Sector	0.556	0.628	0.612	0.645
IRIS	0.501	0.573	0.556	0.584
Gravitational Model				
Centre	0.758			
Sector	0.609			
IRIS	0.543			
Control Case				
Centre	0.541			
Sector	0.498			
IRIS	0.465			

Table 3.28: Comparison of the results of different store selection probability assignment strategies using cross-validation.

We can say from this study that Logit Estimation in which only close large-surface stores are taken into accounnt is easily the most advantageous technique to use. It is the quickest and easiest to use, it relies on the most realistic assumptions about the consideration sets of households, and it produces the most accurate model predictions. This can be improved through the "Hybrid" model, though at a price, as it requires the manipulation of very large data sets.

3.4 Summary

In this chapter, we evaluated the results of our model estimation on the data from our survey, using the Conditional Logit model, using gravitational model, and using the hybrid model of assignment of predicted probabilities.

We compared eight Conditional Logit models, four involving the construction of choice sets containing only close large-surface stores, along with an "outside option", and four involving the sampling of alternatives, either using simple random sampling, or probabilities proportional to size sampling, drawing with respect to probabilities assigned by a gravitational model of store choice.

The estimation of each model's parameters was a time-intensive process that was highly dependent upon the number of alternatives included in the choice set specified for each household. The selection of explanatory variables to include in the model involved a backwards elimination procedure that removed variables that were nonsignificant one at a time until all parameter estimates left had a p-value of no more than 0.15.

The parameter estimates of the conditional logit model were remarkably stable for the different store choices, rarely being of opposite signs. However, the set of variables included in each model were quite different, a reflection of a great deal of redundancy in the variables included in our data set. We found that the parameter estimates of the Conditional Logit models using drawn alternatives were visibly different from the conditional logit models including only near large-surface stores and an "outside option". When we calculated the model predicted probabilities of selection of alternatives associated with each model and calculated the "WD" statistic, we found that defining a choice set in which all far stores were aggregated to form a new alternative produced far superior forecasts of the number of clients visiting each large-surface store than did the predictions of store choice using the sampled alternatives technique, for a comparable

time spent adapting the model. We also found it was far less practical to work with data sets in which there was one predicted probability assigned to every single large-surface store for every household.

In order to provide a simple means of assigning probability estimates to all possible choices of large-surface stores by each household in the sample, we made use of a so-called gravitational model of store attraction. This was in fact a special case of the Conditional Logit model in which the only parameters were the log of the distance of the large-surface store from the household crossed with the store type. We found that our model worked better when we did not enter the retail space of the store as a sort of "mass" term. We could have treated our model as a conditional logit model and used maximum likelihood estimation to estimate the parameters of our model, but we found that the enormous number of large-surface stores in each households' choice set made this very difficult. After a great deal of experimentation, we found that a quick way of obtaining parameters for our model would be to take the cartesian products of all households with all large-surface stores, creating subgroups of these household-store-choice pairs whose associated distance term would have little variability, and then running a regression on the empirical probability of selection associated with each group.

In the last set of models, we used Logit estimation in order to predict the probability of selection associated with the closest stores to each household's domicile, and then used the gravitational model in order to assign probabilities of selection of stores that would be classed as being in the "outside" option. When we tested the "WD" statistic of these models using probabilities of selection calculated on the data set on which the model parameters were adjusted, we found that this provided a very slight improvement in the forecast number of households visiting each store. We conclude by saying that for most purposes an aggregated choice model will work, but for slightly better estimates of

the number of clients visiting each store although at a greater cost, this model can be improved by allocating outside options using a gravitational model of store choice.

Chapter 4

Predicting food product choice through imputation

The first three chapters of this thesis were assigned exclusively to the prediction of large-surface store choice in the Centre Region. However, this choice is only part of a larger pattern of shopping behaviour that was recorded in BVA's survey of shopping flux. Included in this survey were also questions about types of products that were chosen in each store visited, frequencies of store visits, and grocery shopping outside of large-surface stores. We would like to be able to extend our model of store choice so that we can use our survey data to make predictions of these related shopping behaviours.

What we look at in this last chapter of our thesis is how to predict what categories of food products households will select in each of their large-surface store visits. This is a choice set that differs markedly from the choice of large-surface store, in that it has no spatial dimension and we are no longer selecting exactly one of a set of mutually exclusive items, but may select a number of different products in each choice. The food products corresponding to the goods purchased by a given household in a given grocery store is a discrete choice that is too complex to be modeled effectively using classic modeling

techniques such as Conditional Logit, so we wish to use a more empirical method inspired by imputation.

In Section 4.1, we introduce the categories of food products as defined by the data from our survey, and we look at exploratory statistics that describe general patterns of food product choice, how this relates to related behaviours such as frequencies of store visits, and how choices of food products made by the same household within different store choices are correlated. In Section 4.2, we describe various techniques we use for assigning product choices to each household. These are all variations of cold-deck imputation, where individuals whose responses are known are drawn from one data set in order to provide the forecast responses for a data set whose values are unknown. These techniques will then be tested on our survey data in Section 4.3. Because we cannot assume that the choices of large-surface stores will be known when we do prediction, we begin by attempting imputation using only publically available socio-demographic data as the basis for our imputation. However, since the choice of food products must also depend heavily on the stores visited by a given household, we go on to develop an imputation technique that will rely on the predicted probabilities of store selection described in Chapter 3. Our success with this technique will depend upon how reliably we shall be able to use the results of the store choice models in previous chapters.

4.1 Choices of food products

In our survey of shopping behaviour of the Centre Region of France, when asked which store they visited most often, second most often, and third most often, households were also asked how often they visited each store ("several times a week", "once a week", "2-3 times a month", "once a month", or "less than once a month") and were asked which of the following categories of products they selected in each store

- Breads and pastries
- Fruits and vegetables
- Meats and Poultries
- Fish and seafood
- Frozen foods
- Other food-related products

Each household could select any number of these six categories in each of up to three choices of large-surface stores, making a total of 18 decisions per household. What poses a difficulty in modelling this type of multivariate discrete choice is the fact that the different choices of product categories are very correlated. One consequence of this is that we cannot treat the choices of product categories in each of the three large-surface stores of the same household as being independent of each other. In this section, we look at some basic exploratory statistics to look at the interrelations between product choices, and how this relates to some other aspects of the household's choice, such as the type of store in which the products are chosen, the frequency of store visits, and the distance of the store from the households.

We can note some general trends in the choices of product type. In Figure 4.1, we show the percent of households selecting each product type depending on the order of the store choice (provided that a store is chosen). We can see that in general, people will make purchases belonging to more product categories when the store they visit is the store they visit most often than when it is a second or third choice of store. We can also see that people will select products in the "Other" product category in almost every case. This is rather unsurprising, since this product category actually contains most grocery products, including all spices, dairy products, and all canned and pre-packaged goods. As well, we note that people are much less likely to shop for breads and pastries than

any other type of product and are more likely to buy frozen foods, seafood, and fresh food in their first choice of store than in their second and third choices. This may simply reflect the fact that households will buy most of their food stuffs in one store, and then obtain only a few more specialized goods in other stores. The difference between the probabilities of selecting different products in different stores may also be due to the fact that a large number of stores do not offer products in some of the product categories listed above. Not every supermarket includes a bakery, for example, and not all include fresh fish.

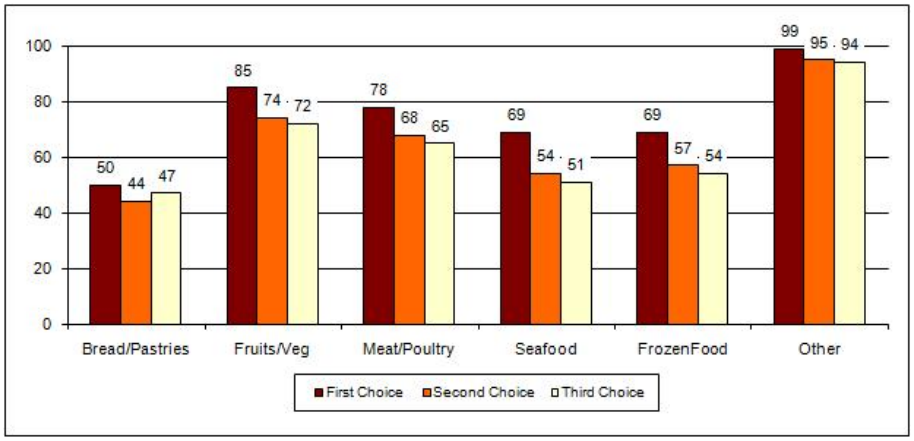

Figure 4.1: Percent of households selecting each product type by order of store choice.

In Figure 4.2, we look at a comparison of the percent of households buying each type of product conditional on the type of store. What is striking here is how close the percent of households selecting each product in supermarkets are to the percent selecting each product in hypermarkets, and how different these percentages are from hard discount stores. The difference can in large part be attributed to the fact that few hard discount stores offer fresh fish and bread, and people may have a lower regard for fresh food at these types of stores.

In Table 4.1, we look at the sets of product categories chosen by each household.

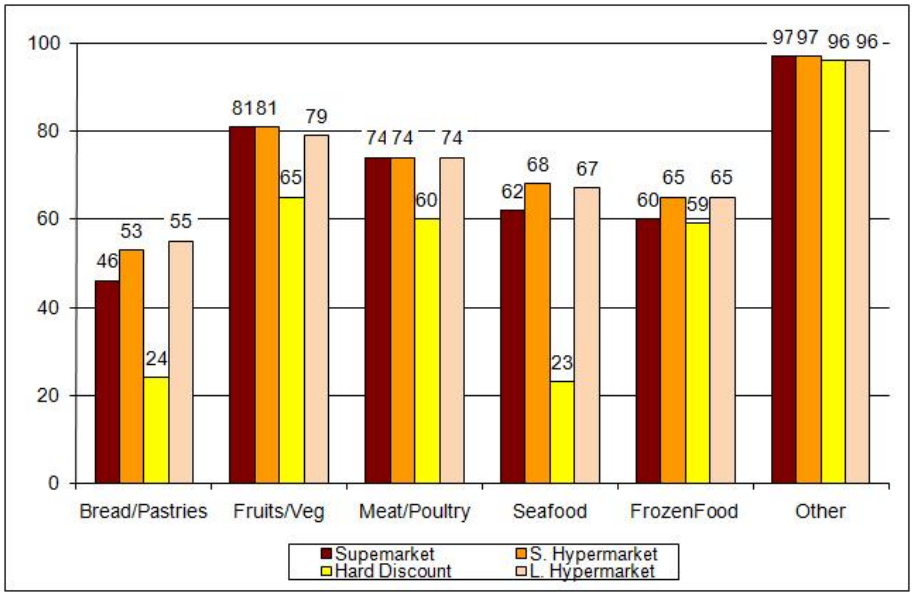

Figure 4.2: Percent of households selecting each product by type of store.

These are indicated with a binary code, where each digit represents the choice of a category. The first digit represents breads and pastries, the second, fruits and vegetables, and so on. So, for example, "011011" represents a household that chose no breads or pastries, that chose fruits and vegetables, that chose meats and deli, that chose no fish or seafood, that chose frozen foods, and that chose products in the "other products" category. We see here that households typically buy either all, or almost all product categories, or if not, they buy only products in the "other products" category. This table shows the percentage of total households in our sample observed selecting each set of products (not all 64 possible combinations of product categories are represented here) in descending order, along with the cumulative percentage.

We see in Figure 4.2 that households buy a greater diversity of products in their first choice of large-surface store than they do in subsequent choices. This phenomenon is borne out in Table 4.2, where we look at the sets of products purchased by each household. The more diverse sets of product categories, like "011111" (all products but

Supermarkets			S. Hypermarkets			Hard Discounts			L. Hypermarkets		
Prod	%	C. %	Prod	%	C. %	Prod	%	C. %	Prod	%	C. %
111111	23.2	23.2	111111	30.0	30.0	00001	13.4	13.4	111111	31.6	31.6
011111	13.0	36.2	011111	12.5	42.5	11011	13.0	26.4	011111	11.3	42.9
011101	6.2	42.4	000001	5.3	47.7	111111	7.8	34.2	000001	6.7	49.6
000001	5.4	47.8	111101	5.2	53.0	010001	7.4	41.5	111101	5.6	55.2
111101	5.0	52.9	011101	5.2	58.1	011001	7.3	48.9	011101	4.6	59.8
011011	5.0	57.9	111011	4.5	62.6	011111	7.3	56.1	111011	4.4	64.2
111011	4.8	62.6	011011	3.6	66.2	111011	5.9	62.1	011011	3.5	67.7
011001	4.5	67.1	010001	2.8	69.0	010011	5.7	67.8	011001	2.5	72.0
010001	3.7	78.0	011001	2.7	71.7	000011	5.4	73.2	111001	2.4	72.6
111001	2.8	73.6	110111	2.1	73.8	001011	5.1	78.2	010001	2.0	74.6
010111	2.3	75.9	111001	2.1	75.9	001001	3.7	81.9	110111	2.0	76.6
010101	2.1	78.0	010111	1.8	77.7	111001	1.9	83.8	010111	2.0	78.6
010011	2.1	80.0	010011	1.7	79.4	001111	1.3	85.1	010101	1.7	83.0
001001	1.5	81.5	010101	1.7	81.1	010111	1.3	86.4	001111	1.5	81.8
110111	1.4	82.9	001111	1.6	82.8	011101	1.1	87.5	010011	1.4	83.3
(B&P|F&V|M&D|F&S|FF|OP)											

Table 4.1: Breakdown of vectors of food product choices by type of store of purchase location.

breads and pastries) are selected much less often in the second and third choice of large-surface stores for the second and third-closest sets of stores, while the less diverse sets of product categories, like "000001" (only products in the "other products" category) are selected more often.

1st Choice			2nd Choice			3rd Choice		
Prod	%	C. %	Prod	%	C. %	Prod	%	C. %
111111	28.4	28.4	111111	22.1	22.1	111111	22.9	22.9
011111	14.1	42.5	011111	14.0	32.5	000001	11.5	34.4
011101	5.8	48.3	000001	9.0	41.4	011111	8.2	42.7
011011	5.5	53.8	011011	4.8	46.3	111101	4.4	47.1
111011	5.4	59.2	011101	4.5	50.8	010001	4.3	51.4
111101	5.4	64.6	011001	4.4	55.2	011001	4.2	55.6
011001	3.4	68.0	010001	4.1	59.2	011101	4.2	59.8
000001	3.0	71.0	111011	4.1	63.3	011011	4.1	63.9
010001	2.8	73.8	111101	4.1	67.4	111011	4.1	68.0
010111	2.5	76.3	010011	2.6	70.0	111001	2.8	70.9
111001	2.3	78.6	111001	2.5	72.5	000011	2.2	73.0
010101	2.1	80.7	001001	2.0	74.5	001001	2.2	75.2
010011	2.0	82.7	000011	1.9	76.3	010011	1.8	77.0
110111	1.9	84.6	001011	1.6	77.9	110001	1.5	78.5
001111	1.8	86.4	010101	1.5	79.5	001011	1.4	80.0
(B&P|F&V|M&D|F&S|FF|OP)								

Table 4.2: Breakdown of vectors of food product choices by order of store choice.

This shows us that choices of products are clearly not independent, since people are more likely to buy some products if others are bought as well. The Pearson Correlation Coefficients between the empirical likelihoods of selecting each of these products for a single store visit are in Table 4.3. All of them are significant, showing a strong correlation between all choices of products by households in the population

The strongest correlations are found between Fruits and Vegetables and Meats and Deli products. These are both categories of produce that may seen as "fresh food" by

Category	B&P	F&V	M&D	F&S	FF	OP
Bread/Pastries	1.00	0.25	0.26	0.24	0.24	0.03
Fruits/Vegetables	0.25	1.00	0.38	0.30	0.28	0.10
Meats/Deli	0.26	0.38	1.00	0.34	0.31	0.09
Fish/Seafood	0.24	0.30	0.34	1.00	0.29	0.08
Frozen Foods	0.24	0.28	0.30	0.29	1.00	0.12
Other Products	0.03	0.10	0.09	0.08	0.12	1.00

Table 4.3: Correlation between choices of products between large-surface store visits undertaken by the same household.

consumers, and thus it may be that if a consumer will be willing to select one of these categories in one store, it will be willing to select the other. The weakest correlation is between each product category and "Other Products". People seem to buy in the "Other Products" category just about every time, regardless of what other products they buy in the store.

The frequency of store visits to each store is also a variable that we hope to predict in subsequent work. In Table 4.4, we look at the relationship between the observed store frequencies (in average number of store visits per week) and the empirical probability of selecting a given product. We do see linear relationships between the frequency of store visits and the range of products households will buy. In all categories but "Other Products", the empirical probability of selecting a given product goes up by at least 20 percent as we go from few store visits to many store visits.

The distance of a given store from a household's domicile is the factor we see as most important in determining its probability of being selected by the given household. Surprisingly, this distance produces no change in the type of products bought in the given store, as is seen in Table 4.5.

We have noticed a slight relationship between type of store and order of store choice and the probability of selecting each type of product. However, we find that a much more

	Visits/week				
Food categories	1/10	1/4	2/5	1	4
Bread/Pastries	0.38	0.42	0.44	0.48	0.59
fruits/Vegetables	0.65	0.69	0.76	0.84	0.88
Meats/Deli	0.57	0.65	0.70	0.77	0.80
Fish/Seafood	0.50	0.52	0.57	0.65	0.70
Frozen Foods	0.47	0.54	0.59	0.67	0.70
Other Products	0.95	0.95	0.96	0.98	0.98

Table 4.4: Empirical probability of selecting product type by frequency of store visits.

	Dist in km							
Food categories	0.50	0.75	1.50	3.00	7.00	15	25	50
Bread/Pastries	0.49	0.48	0.46	0.48	0.49	0.46	0.45	0.49
fruits/Vegetables	0.81	0.78	0.78	0.78	0.80	0.80	0.77	0.78
Meats/Deli	0.73	0.71	0.71	0.72	0.73	0.74	0.72	0.71
Fish/Seafood	0.59	0.59	0.61	0.60	0.62	0.64	0.60	0.59
Frozen Foods	0.63	0.62	0.63	0.64	0.64	0.60	0.60	0.58
Other Products	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.96

Table 4.5: Empirical probability of selecting product type by distance of store location in km.

important phenomenon that we observe is the correlations between the probabilities of selecting each product in different choices of large-surface store made by the same individual. We find that when a household chooses one type of product in one of its choices of stores, it will tend to choose this same product in the other choices of stores.

We have here calculated the empirical probability of purchasing each product category conditional on the product category chosen in an anterior store choice. We illustrate this in the simple bar graph in Figure 4.3. The first two bars compare the probability that a household will choose breads and pastries in its second choice given that it doesn't in its first choice, and the probability that a household will do so if it did choose breads and pastries in its first choice. Similarily, for the other two pairs of bars, we show the probability of selecting breads and pastries in the third choices given the first choice, and the probability of selecting breads and pastries in the third choice given the second choice. The conditional percentages for all the products are listed in Table 4.6. As we can see here, the values in the second column for every single category of food product and every condition is much higher than the first column, meaning that households have a tendency of buying the same products in second and third choices of large-surface stores that they chose in the anterior store choices. The dependence of the probabilities seems to be strongest for frozen foods, and least strong for seafood. This seems to imply that the choice of frozen foods is more dependent on the person, while the choice of seafood is more dependent on the choice of store. People may not tend to buy seafood in more than one place as much as they would other products, while the purchase of frozen foods may be more a matter of personal preference than a matter of opportunity.

In Table 4.7, we examine behavioural continuity again, this time looking at household choices of entire sets of product categories, looking at the empirical probability that a household will select the same set of products in a subsequent choice of large-surface

Figure 4.3: Probability of selecting Breads and Pastries conditional on previous choices.

Product	Condition	Noncontinuity	Continuity
Bread/Pastries	Ch 2 Given Ch 1	0.23	0.64
	Ch 3 Given Ch 2	0.27	0.68
	Ch 3 Given Ch 1	0.26	0.63
Fruits/Vegetables	Ch 2 Given Ch 1	0.37	0.80
	Ch 3 Given Ch 2	0.43	0.79
	Ch 3 Given Ch 1	0.39	0.76
Meats/Deli	Ch 2 Given Ch 1	0.30	0.77
	Ch 3 Given Ch 2	0.35	0.78
	Ch 3 Given Ch 1	0.29	0.73
Fish/Seafood	Ch 2 Given Ch 1	0.34	0.63
	Ch 3 Given Ch 2	0.30	0.66
	Ch 3 Given Ch 1	0.34	0.58
Frozen Foods	Ch 2 Given Ch 1	0.21	0.71
	Ch 3 Given Ch 2	0.23	0.73
	Ch 3 Given Ch 1	0.22	0.65

Table 4.6: Conditional percent of households choosing each product based on anterior choices

store that it chose in an anterior store choice. Here, "2G1" means "Second choice given first choice". Thus, the first value in Table 4.7 is the empirical probability, conditional on there being a first and a second choice of large-surface store, that the second choice of product categories will be the same as the first, if the first set of product categories is labeled "111111", meaning all product categories selected.

Vector	2G1	3G2	3G1
111111	19.9	66.1	52.8
011111	23.0	70.1	60.7
111101	26.5	68.9	56.8
011101	31.3	75.1	62.9
111011	18.0	62.7	52.5
011011	17.4	64.8	56.5
011001	25.2	75.8	62.9
010001	37.0	76.0	62.9
001001	30.3	80.7	62.3
000011	35.1	76.2	63.7
000001	45.3	87.1	71.8

Table 4.7: Continuity of choices for all product choices combined.

From looking at this data, we can see that there is a slight relationship between order of store choice, frequency of store selection, store type and products chosen, but this is not very strong. However, we do find that the choice of products tends to depend upon the individual decision maker, and are correlated. People tend to make the same choices of products for their few choices of stores, and product choices tend to be clumped together, some people choosing most product categories, and some choosing only products from the "Other Products" category.

The advantage of using imputation for prediction is that we can preserve the correlation between product choices made for different store choices made by the same household without requiring us to understand and specify this. Since our choices do not

depend on spatial characteristics, and only in a limited way on the actual store choice, we hope to be able to capture the effect of individual preferences through sociodemographic variables.

4.2 Prediction through imputation

In order to predict the choice of products for each household's three choices of stores (or nonchoices), we have a wealth of explanatory variables from our survey data referring to the characteristics of the household and the geographic characteristics of the household's neighbourhood. We decide to assign product choices based on cold-deck imputation. In this section, we shall discuss why we choose this technique and what this technique entails. We then shall introduce the variations of this technique: imputation by class defined by crossed auxiliary variables, imputation by class defined by model scores, and imputation by nearest neighbour.

Since we have the same finite set of product categories for every choice of large-surface store for every household, we may be tempted to use Logit estimation as we did for store choice. However, we believe that this is not very well-adapted to our purposes. Logit models are used to assign probabilities of selection to discrete choices that have a certain defined structure. Individuals choose one and only one alternative, and the choice probabilities must satisfy the independence of irrelevant alternatives property; that is, that ratios of any two probabilities of selection of any alternative must remain independent of the ratios of any two other alternatives. The problem is that since households may select more than one product, these product categories do not serve as well-defined alternatives for the Logit model. In fact, the set of mutually exhaustive and mutually exclusive alternatives selected by each household in each of the stores it visits is the set of 2^6 combinations of category choices possible. The calculation time of the estimation

of the parameters of the Conditional Logit model is very sensitive to the number of alternatives in each choice set, especially in the case of constant explanatory variables that require a different parameter for each alternative. A model with 64 alternatives, such as our model, is very cumbersome to run. The independence of irrelevant alternatives is also not verified in the case of choices of product categories, since different alternatives will contain overlapping choices of products. For example, the ratio of the probability of selecting the category of "meats and poultries" and the probability of selecting both the categories "meats and poultries, and "fruits and vegetables" will not be independent of the ratio of the probability of selecting the category "breads and pastries" and the probability of selecting both the categories "breads and pastries" and the category "fruits and vegetables", since both these ratios will diminish if the inherent attractivity of the category "fruits and vegetables" for a given household increases. If we insist on using a model based on Logit, perhaps by using an extension of Logit proposed by Hendel (1999, [19]) that allows multiple choices per individual, it would be difficult to incorporate all the correlations that exist between the choice of each food category.

We therefore believe that an expeditious manner of making a prediction for a given household, based on recorded behaviour, would be not to use any model or regression techniques, but to treat the choices of products that we wish to predict as if they were missing values that we impute with values from a data set of recorded household choices. Imputation refers to the replacement or attribution of values to missing or erroneous data entries in an incomplete data set. The value that is imputed is based upon the data available to the statistician. It could come from either the correct data entries within the same data set, or perhaps from an external data set. A typical use of imputation is the assignment of responses to nonrespondents in a survey, based upon the sociodemographic characteristics of the respondent. Due to the widespread use of imputation in order to

deal with inevitable data collection problems, this is a well-developed field and many techniques exist today. Some of these techniques include: Imputation by Regression, Imputation by Regression with an added Residual, Imputation by Previous Value (in longitudinal data), Imputation by Nearest Neighbour, and Deck Imputation. These different techniques are discussed in Haziza (2002 [17]).

Deck imputation refers to the replacement of a missing or erroneous observation (the "receiver" observation) by another observation (the "donor" observation) that is chosen from either within the same data set ("hot-deck" imputation) or from another data set ("cold-deck" imputation). The technique we will be using is cold deck imputation. Normally, this term refers to imputation done using random draws of households from another data set, but we use this term in a larger sense of simply choosing a household from the "cold deck" to match the receiver household, whether using random sampling or not. In our case, the "cold deck" will be the data from BVA's survey of household shopping behaviour, that we shall use to provide donor households to any data set of households whose choices of food products we wish to predict.

The quality of the cold-deck imputation will depend upon the likelihood that donor responses are equal to the true receiver responses to which they are assigned. We suppose that this likelihood is greater the more the donor individuals resemble the receiver individuals. We therefore make use of a set of variables, called the auxiliary variables, that are known for both the donor and receiver households in order to determine the donor individual that most resembles the receiver. Sections 4.2.1 to 4.2.3 describe three methods that we use in order to do this.

4.2.1 Imputation by class defined by crossing auxiliary variables

If the auxiliary variables that we wish to integrate into our imputation are categorical, then the combinations of modalities observed for each donor and receiver household can be used to define classes to which they can be assigned. To predict any given response, we simply determine the imputation class of the individual, based on its characteristics, and then draw an observation at random from the same imputation class in the donor sample. For example, if we were to define our imputation class by the employment and household type of a household, then if we wished to predict the products chosen for a household headed by a couple in which both members work, we would draw a household at random from only those households in our sample from the Centre Region headed by couples in which both members work. The predicted choice of products for the three choices of large-surface stores for this first household will simply be the choice of products recorded for the drawn household. We can extend this method to predict/impute the products chosen by all the households in an entire population. When we do this, we begin with a prediction population, for whom we do not know the products chosen, and a donor, or reference population, for which we do, both of which are divided into imputation classes. For every household in the prediction set, we do a separate, independent draw from the same class in the reference set to obtain the individual who provides the imputed values.

We illustrate this in Figure 4.4. Here, we portray each household in the donor and the receiver samples as occupying a point within the space of auxiliary variables "X1" and "X2". The dark points are the receiver observations, and the light points are the donor observations. On the left, we see the two samples before imputation. The responses, 1, 2, or 3 of the donor respondents are known, and the responses of the receivers, the households whose responses we wish to predict, are unknown. The two sets of observations are divided into subsets according to the values of the auxiliary variables "X1" and "X2".

Each receiver household is then paired with a donor household, and the donor response is imputed as the receiver response, as shown on the right. We see that we do not need to have the same number of receivers as donors in each subset, but each receiver household must be in an imputation class that includes at least one donor observation for it to be assigned a response value.

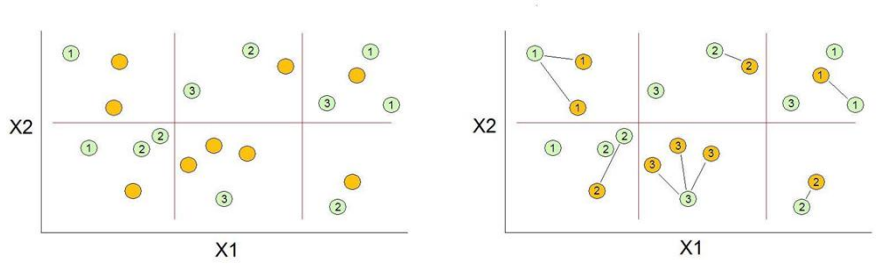

Figure 4.4: Imputation by imputation class.

Haziza (2002 [17]), discusses the advantages and disadvantages of several forms of imputation before discussing imputation using classes of imputation defined by taking the cartesian product of qualitative variables, described here. He suggests that in order to use this technique, it would be useful to use modelling procedures to determine which auxiliary variables are correlated with the response variable. If the variables are well-chosen, the response variable will have little variation within the imputation classes.

One problem with this technique, however, is that it takes few auxiliary variables to obtain an enormous number of imputation classes, and if we use all significant variables, we will have many imputation classes for which we have no observed response, and so no imputation can be made. Even if we restrict ourselves to using only a few variables, we are still likely to have a problem, since there will always exist "corner" and "edge" cases of rare combinations of auxiliary variable modalities. There are few viable answers to this problem.

Haziza has suggested that variables be ordered in importance in terms of their corre-

lation with the response variable, as determined by previous modeling work. This way, imputation classes can be constructed by crossing variables one at the time, following the order that has been established. The process is an iterative algorithm:

1. We begin with imputation classes constructed using an initial set of auxiliary variables that we have considered to be the most important.

2. We cross this set of imputation classes with the most important auxiliary variable not taken into account in the determination of the imputation classes, in order to create a new set of imputation classes.

3. We then examine these new imputation classes in order to see which do not conform to two preset criteria:

a. The minimum number of nonmissing responses is below a preset value K.

b. The number of missing responses is greater than the number of nonmissing responses.

4. The new imputation classes that do not meet these criteria are then aggregated with other new imputation classes that are found within the same old imputation class, until all the members of our new set of imputation classes meet the above criteria.

5. We go back to step 1, proceeding to the next most important explanatory variable, or ending our process of creating imputation classes when it is evident that no new auxiliary variables can be taken into account.

In our case, we would ignore the second of the two criteria, since we must be able to impute the values of millions of missing product categories using only a few thousand observed values. Haziza has acknowledged that aggregating imputation classes will introduce bias into the model, and prefers other techniques of creating imputation classes such as the "method of scores".

4.2.2 Imputation by class defined by model scores

The problem with the previous technique is that by defining an imputation class as being all the combinations a set of auxiliary variables, one is likely to have a vast number of redundant classes. We can see this in a very simple example. Imagine that we wish to predict whether a household makes use of a monthly metro pass or not. We can make a distinction between households who regularly use a car, and those who don't. If we had a second variable indicating if a household lived near a metro station or not, it may not make sense simply to cross the two variables to make four groups. There may be a meaningful distinction between people who don't use a car who live near the metro (and therefore would be more likely to use it regularly) and those who don't. However, if it is the case that households who drive don't use public transportation, then there is no reason to make a distinction between those living near the metro and those who do not.

The method of scores allows us to find those combinations of auxiliary variable values that define important distinctions between behaviour patterns and those that don't. Instead of looking at the auxiliary variable values directly, we use them in a modelling procedure that produces one or more scores. We could use Logistic Regression, for example, to predict the probability of a variable related to the response. This score is then used to define imputation classes. The advantage of this method, is that it can reduce the number of auxiliary variables used without sacrificing too much relevant information. Continuous variables can also be used to help define the score. Once the scores are defined, we can use them to define imputation classes, either by using classification procedures, or by ranking the scores and dividing the population into equal-sized subsets according to this rank.

4.2.3 Nearest Neighbour Imputation

If our auxiliary variables are continuous, an alternative to imputation by imputation class consists of defining a distance measure on the auxiliary variables of our sample, and then using this to determine the donor observation closest to the receiver observation in the auxiliary variable spaces. This is normally a deterministic operation that always yields the same imputed values.

An example of nearest-neighbour imputation might be a case where we wish to impute the value of a household's Christmas-season spending, and we have only one auxiliary variable: annual revenue. In that case, we would impute the Christmas-season spending observed for the household in the donor set having an annual revenue that was the closest to that for the household in the receiver set.

We can illustrate this in the following diagram:

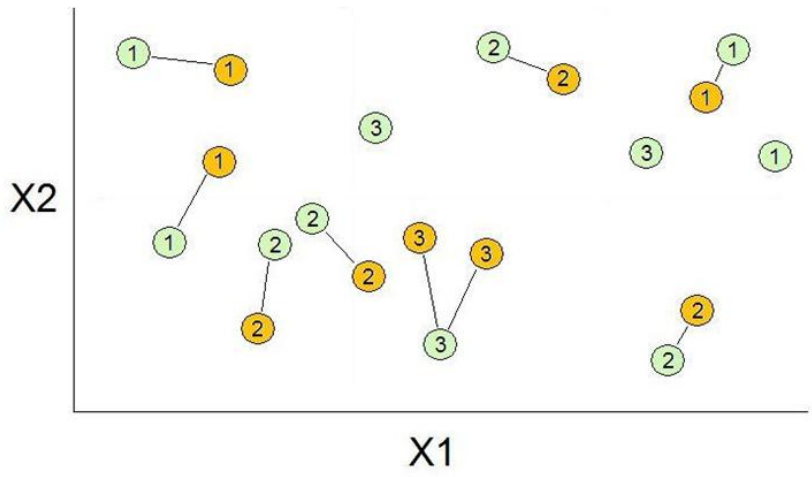

Figure 4.5: Imputation by Nearest Neighbour.

Once again, the donor observations are light, and the receiver observations are dark. However, the auxiliary variables "X1" and "X2" are continuous here and the diagram

shows the donor observations paired with the closest receiver observations. An advantage of this method is that the problem of "empty classes" with no donor responses is avoided since every single individual is guaranteed to have a donor household. The degree to which this method is valid is the degree to which the ranges of the receiver and donor observations overlap in the space defined by the auxiliary variables. A disadvantage of this approach is that the determination of the nearest neighbour to an individual can be time-consuming since it demands the calculation of many distance values.

4.2.4 Evaluating forecasts by imputation

In order to test the validity of our imputation, we use cross-validation by department of residence in the same sort of way as what we described in 2.2.3. We do imputation separately for each department of our survey sample. At each stage, one department in the sample defines a receiver sample, and all other departments in the survey become donor households. Once imputation is done on each department, we will have assigned a predicted choice of products to each household selected for our survey, and we can compare these values to those observed for these same households. If we refer again to our illustration in Figure 2.3, the dark departments this time represent the donor samples whose values are used to impute the unknown product choices in the receiver department, which are now the light departments. The donor set in imputation corresponds to the training set in our Logit models, and the receiver set in imputation corresponds to the test set.

For each product choice, we calculate the percentage of cases where the imputed response matched the true response. If there were N households in the survey sample, and three responses per household, for each product, there were $3N$ cases. The formula for this statistic, we shall call "WA" for well-allocated percentage. The formula for

calculating the "WA" statistic for product q is:

$$\text{WA}(q) \quad = \quad \sum_{k=1}^{3} \sum_{n \in N} \frac{1 - |O_{nkq} - A_{nkq}|}{3N}$$

where N refers to both the set of households in the survey sample, and the number of individuals in the sample. O_{nkq} is one when household n selects product q in its k^{th} choice of store and zero otherwise. Similarly, A_{nkq} is one when household n is assigned product q (by imputation) in its k^{th} choice of store and zero, otherwise.

One problem with this method is that by calculating the marginal probabilities of each product type being well-allocated, we neglect to measure the accuracy of the interactions between the forecasted choices of products. We have therefore also calculated the percent of cases where all six product choices were well-allocated.

In order to have an evaluation of the assignment of product choices overall for each household, we also calculate a sum of squares of the difference between the predicted and observed number of households selecting each combination of store choices in each survey sector. Let c represents a single one of the 64 combination of choices of product types for one store visit. We define O_{nkc} as one if household n selects this combination of product types for choice k of large-surface store, and zero otherwise. Similarly, A_{nkc} is one if n is assigned c for choice k of large-surface store and zero, otherwise. This means that the observed and the expected number of store visits corresponding to a combination of store products choices will be, respectively:

$$O_c \quad = \quad \sum_{n \in N} \sum_{k=1}^{3} O_{nkc}$$

and

$$E_c = \sum_{n \in N} \sum_{k=1}^{3} A_{nkc}$$

We can use these values to calculate the statistic that we call "SSD" for Sum of Squared Differences:

$$\text{SSD} = \sum_{c \in K} \frac{(O_c - E_c)^2}{O_c}$$

If we treat the imputed responses for each household as random values, then SSD corresponds to the famous "Chi-squared Test" of the null hypothesis that the expected value of E_c is equal to O_c and that the two values are independent. If we wish to develop an imputation technique that will generate the same responses as those observed for each individual, than we wish SSD to be as low as possible. We can compare the values of SSD with the some of the p-values of the chi-squared distribution:

-P(SSD $< 40) < 0.01$
-P(SSD $< 46) < 0.05$
-P(SSD $< 50) < 0.10$
-P(SSD $< 54) < 0.20$
-P(SSD $> 63) < 0.50$
-P(SSD $> 73) < 0.20$
-P(SSD $> 78) < 0.10$
-P(SSD $> 83) < 0.05$
-P(SSD $> 93) < 0.01$

Because it was not only the assignment of store choices in the entire Centre Region that interested us, but the assignment of choices at the level of survey sectors, we also create the SSD(S) statistic that looks at the difference between the observed and

predicted product categories broken down by sector. This formula becomes

$$\text{SSD(S)} = \sum_{s} \sum_{c \in K} \frac{(O_{sc} - E_{sc})^2}{O_{sc}}$$

with

$$O_{sc} = \sum_{n \in s} \sum_{k=1}^{3} O_{nkc}$$

and

$$E_{sc} = \sum_{n \in s} \sum_{k=1}^{3} A_{nkc}$$

where s represents the survey sector. We have to eliminate any terms in which O_{sc} is zero, leaving us with a "Chi-squared Statistic" with 6103 degrees of freedom. Unfortunately due to the large number of terms in SSD(S) that are sums of very few observations, we cannot assume that SSD(S) follows a Chi-squared distribution. However, as a guideline, if SSD(S) did follow a Chi-squared Statistic, then

-P(SSD(S) $>$ 6288) $<$ 0.20
-P(SSD(S) $>$ 6246) $<$ 0.10
-P(SSD(S) $>$ 6198) $<$ 0.05
-P(SSD(S) $>$ 6105) $<$ 0.01

4.3 Application to survey data

In this chapter, we apply the three imputation techniques described in Section 4.2 to our survey data. Obviously a household is not going to buy a food product in a store

that it does not visit, so whether or not a household makes a purchase in a particular product category will depend a great deal on whether the household visits a store or not. Unfortunately, this information is not known to us in advance. For this reason, we begin by attempting to assign choices of food product categories solely on the known socio-demographic characteristics of each household and show that these variables have very little effect on the accuracy of the household's predicted choice. We then look at whether we can develop a manner of predicting choices of food product categories based on the predictions of the store choice model we developed in the first three chapters of this thesis.

4.3.1 Imputations based on socio-demographic variables

The complete list of all the socio-demographic variables that we have at our disposal as possible auxiliary variables in our imputation is in Glossary 2 at the end of the thesis. We used these variables to define imputation classes in two ways: first by crossing the variables so that each combination of each of the variables' modalities defines an imputation class, and secondly by using a Logistic model to assign a set of scores that are used in a clustering procedure in which the resulting clusters are taken as imputation classes. We shall look at the results of both of these techniques.

Crossing auxiliary variables

In order to determine the sociodemographic variables that have the greatest effect on the probabilities of selecting each product type, we use a stepwise method of variable selection in 18 Logistic Models, each one used to predict the probability that a household selects a different combination of product category and order of store choice. The actual model estimates are not important, since they are merely used to determine in how many

of these Logistic models each variable is significant (at a p-value of 0.05). Table 4.8 shows the most important sociodemographic variable along with the number of times each was found to be significant.

Variable	# Significant
agecat	12
Carcat	9
revcat	8
cplcat	7
Csupcat	5
vehcat	5
PropLocaCat	4
UCCat	4
Hsizecat	3
PropAP82Cat	3
PropRetrCat	3
achicat	3
piedcat	3
rescat	3
transcat2	3
tu299n	3
tu99n	3
AccVilFCat	2
NonAl	2
PropTertCat	2
anlgcat	2
avgnachiCat	2
medUCCat	2

Table 4.8: Sociodemographic variables that contributed most to explaining the choice of food categories.

We use these variables to define choice sets by adding one variable at a time to the set of variables that define our imputation classes. We define these imputation classes in Table 4.9. In Class A, there is only one imputation class that includes the entire population, so when we do cross-validation, the imputed choices of food products are taken from a household chosen at random from the entire donor data set. In Class B, we restrict ourselves to only selecting a donor household for imputation that is in the same

A	No auxiliary variables
B	agecat
C	agecat, carcat
D	agecat, carcat, revcat
E	agecat, carcat, revcat, cplcat
F	agecat, carcat, revcat, cplcat, csupcat
G	agecat, carcat, revcat, cplcat, vehcat

Table 4.9: Imputation classes defined by categorical variables

age category as the receiver household. In Class C, we only select a household that is in the same age category, and that has the same value for the carcat variable (the variable specifying whether members of the household commute regularily by vehicle).

Class	B&P	F&V	M&D	F&S	FF	OP	Vect	SSD	SSD(S)	# Imp
A	0.61	0.63	0.61	0.61	0.61	0.71	0.24	78.7	14161	12573
B	0.61	0.64	0.62	0.61	0.62	0.72	0.25	90.4	14365	12573
C	0.61	0.64	0.62	0.61	0.62	0.72	0.25	62.2	14593	12573
D	0.62	0.64	0.62	0.61	0.63	0.72	0.26	99.0	14430	12573
E	0.61	0.64	0.62	0.62	0.62	0.72	0.25	83.6	14531	12562
F	0.62	0.64	0.62	0.61	0.62	0.72	0.26	75.9	14097	12538
G	0.62	0.64	0.62	0.61	0.62	0.72	0.26	81.0	14151	12527

Table 4.10: Imputation using classes defined by crossed variables

With these variables, we go on to do cross-validation on the Centre Region generating the statistics that we introduced in Section 4.2.4. These are presented in Table 4.10, which shows the evaluative statistics associated with the imputation classes defined in Table 4.9. The first six numbers are the well-allocated statistics calculated for each product category. We follow that with the percent of times that every choice of product category is well-predicted for a household's choice of store. The two columns before the last represent the sum of squared differences between the observed and predicted number of households selecting each combination of product categories over the entire region, and

tbph

Variable	# Significant
agecat_4	12
Carcat	9
agecat_3	8
agecat_2	7
propcat	6
UCCat_2	5
Csupcat	4
revcat_2	4
revcat_4	4
PropAP75Cat_1	3
achicat_3	3
anlgcat_2	3
piedcat	3
quotacat_3	3
vehcat	3

Table 4.11: Dummy variables of sociodemographic variables that contributed most to explaining the choice of food categories.

the sum of squared differences for each combination of product categories for each survey sector. The last column is a count of the number of households in the Centre Region for whom we were able to select a donor household for imputation. As we can see, as we increase the number of variables, there will be more and more possible combinations of values of explanatory variables, and more chance that there some combinations will be present in a receiver household, but in no donor households. From looking at Table 4.13 the use of these imputation classes is not more effective at allocating choice probabilities than a simply drawing household at random, and becomes much worst when we start using more auxiliary variables to define imputation classes.

In order to ensure that our choice of variables are not distorted by the fact that different categorical variables have different numbers of modalities, we also look at the binary dummy variables that represent each categorical variable. Once again, using Logistic Regression, we look at the variables that are significant in the largest number of

A	No auxiliary variables
B	agecat_4
C	agecat_4, Carcat
D	agecat_4, Carcat, agecat_3
E	agecat_4, Carcat, agecat_3, agecat_2
F	agecat_4, Carcat, agecat_3, agecat_2, propcat
G	agecat_4, Carcat, agecat_3, agecat_2, propcat, UCCat_2
H	agecat_4, Carcat, agecat_3, agecat_2, propcat, UCCat_2, Csupcat
I	agecat_4, Carcat, agecat_3, agecat_2, propcat, UCCat_2, revcat_2
J	agecat_4, Carcat, agecat_3, agecat_2, propcat, UCCat_2, revcat_4
K	agecat_4, Carcat, agecat_3, agecat_2, propcat, UCCat_2, revcat_4, csupcat
L	agecat_4, Carcat, agecat_3, agecat_2, propcat, UCCat_2, revcat_4, revcat_2, csupcat

Table 4.12: Binary auxiliary variables used to define imputation classes

	B&P	F&V	M&D	F&S	FF	OP	Vect	SSD	SSD(S)	# Imp
A	0.60	0.63	0.61	0.60	0.60	0.71	0.24	83.6	14941	12573
B	0.61	0.64	0.62	0.61	0.62	0.72	0.25	78.6	14436	12573
C	0.62	0.64	0.62	0.62	0.62	0.72	0.25	83.6	14441	12573
D	0.62	0.64	0.63	0.61	0.62	0.72	0.25	91.8	15548	12573
E	0.62	0.64	0.62	0.61	0.62	0.72	0.25	65.3	14678	12573
F	0.62	0.64	0.62	0.61	0.62	0.72	0.25	70.8	14506	12573
G	0.62	0.64	0.62	0.61	0.62	0.72	0.25	58.4	14704	12573
H	0.62	0.64	0.62	0.61	0.63	0.72	0.25	56.5	13982	12573
I	0.62	0.64	0.62	0.61	0.63	0.72	0.26	68.0	14767	12573
J	0.62	0.64	0.62	0.61	0.62	0.72	0.25	76.5	13514	12573
K	0.62	0.64	0.62	0.61	0.63	0.72	0.25	56.5	13982	12573
L	0.62	0.64	0.62	0.61	0.63	0.72	0.26	68.0	14767	12573
M	0.62	0.64	0.62	0.61	0.62	0.72	0.25	74.4	14700	12572

Table 4.13: Imputation done by class defined by crossing binary dummy variables

Clusters	B&P	F&V	M&D	F&S	FF	OP	Vect	SSD	SSD(S)	# Imp
1	0.60	0.63	0.61	0.61	0.60	0.71	0.24	53.7	14531	12573
5	0.62	0.64	0.62	0.61	0.62	0.72	0.25	61.2	13928	12573
10	0.62	0.63	0.62	0.61	0.63	0.72	0.25	67.3	14073	12573
20	0.62	0.64	0.62	0.61	0.62	0.72	0.25	68.5	14114	12573
50	0.62	0.64	0.62	0.61	0.62	0.72	0.25	66.7	14195	12573

Table 4.14: Imputation done by class defined by clusters of vectors of predicted probabilities

choices of product categories as possible. These are shown in Table 4.11. In Table 4.12 we show the combinations of these variables that generate the imputation classes which we test with cross-validation in Table 4.13. The conclusion we draw from this table are identical to those drawn from Table 4.10. As we can see once again in Table 4.13 by the fact that the values of the first six columns remain constant, and that every value of SSD is over 54, our imputation classes certainly don't improve our predictions of choices of choice of food category.

Using clustering of scores

Due to the problem with crossing large numbers of variables in order to generate imputation classes, we try to modify our technique. Instead of looking at all the combinations of values of all the explanatory variables that interest us, we generate our imputation classes through the calculation of a score. The way we apply this technique to our data is to use the parameters calculated for the Logistic Regressions undertaken on all donor samples, and use these in order to assign probabilities of selection to every choice of product category for each choice of large-surface store made by each individual. These predicted probabilities over the entire sample of both donor and receiver households is then divided into a number of imputation classes using a classification procedure in SAS

called FASTCLUS. Donor household are drawn for each receiver household from the same FASTCLUS cluster. As we can see in Table 4.14, for various numbers of clusters, the results of the imputation remain essentially the same, although here, we can be assured that all the information contained in the explanatory variables has been taken into account, and we have no problem assigning donor individuals to individuals whose response we wish to impute. We note that in Table 4.10, Table 4.13, and Table 4.14, the first line of the table corresponds to the same imputation technique where the donor household is drawn without any regards to its characteristics. The reason for the difference in the SSD and SSD(S) terms in these three tables is simply the random flux introduced due to the fact that we do a separate draw for each one. This should give some perspective in attempting to interpret "improvements" in the SSD term.

Although in this section, we are able to introduce different imputation procedures, and are able to take into consideration an enormous number of variables using different techniques, we are forced to conclude that household characteristics do not sufficiently determine choices of food products for our purposes.

4.3.2 Imputation based on characteristics of store choice

In this section, we will look at how we can take into consideration the choice of large-surface store made by each household in predicting the associated choices of food products. We begin by doing imputation by class, this time using as auxiliary variables the characteristics of the choices of large-surface stores made by each household.

When we apply this technique to our case at hand, we have two options. When we wish to impute the missing product categories for a given household, we can either draw a different individual from the imputation class for each choice of large-surface store, imputing the products associated with each choice independently, or we can restrict

ourselves to imputing the missing product categories associated with all three of the store choices of a given individual using the values observed for only one individual drawn from the imputation class.

In the first case, we treat the three choices of large-surface stores assigned to each household as independent units. In the donor sample, these household-store-choice pairs are then divided into imputation classes according to the characteristics of the associated household, and according to the characteristics of the store choice. When we wish to impute the product categories associated with a given households' three choices of stores, then we associate each of the household's choices with a household-store-choice pair in the reference class from which we draw one pair whose values will be used for imputation. We note that the product categories imputed for each of the three choices of large-surface store for a given individual do not necessarily come from the same individual in the donor set. For example, suppose that we have assigned a household a supermarket in a suburban commune for a given individual's first choice of store, a large-hypermarket in a downtown commune for the individual's second choice of store, and no store for the household's third choice of store. Suppose our hypothetical individual is a retiree and that the imputation classes are defined by age category, by store type, and by commune type. If we wish to impute the missing product categories associated with each of the household's store choices, we would begin by drawing a retiree observed visiting a supermarket in a suburban commune at random from the reference population. Suppose we find a retiree in the reference population who selects such a supermarket as its second choice of large-surface store (since in this case, the order of store choice is not used in the definition of the imputation classes). If he/she bought products in the fruits-and-vegetables and meats-and-deli categories, these categories would be imputed. For the next choice of store to be imputed, we draw a retiree observed visiting a large hypermarket in a downtown

commune. This could be a retiree who selects such a hypermarket as its second choice of store, in which case, it would obviously not be the same retiree as the one selected for the first choice of store. If this drawn retiree selects products from all product categories, then the predicted shopping behaviour of our individual will be a choice of fruits-and-vegetables and meats-and-deli products chosen in a supermarket in a suburban commune, all product categories purchased in a downtown large hypermarket, and no store visit as its third choice.

In the second case, we would select one individual from whom we would obtain the values to be imputed for all three choices of large-surface store. If we return to the same hypothetical retiree, we could define our imputation class in terms of age, and store types chosen. Thus, we would draw a retiree visiting a supermarket as his/her first choice of store, and visiting a large-hypermarket as his/her second choice of store. If this retiree bought only fruits and vegetables in the supermarket and only frozen foods in the hypermarket, then this will be the predicted behaviour of our individual.

An advantage of the first technique is that the characteristics of a triplet of three store choices relevant to predicting shopping behaviour is more difficult to find than the characteristics of a single choice. However, the implicit assumption of this technique is that the products chosen by the individual for the three choices of stores are independent, something that is known to be false.

Before we can take into account the characteristics of store choices, we need to assign three choices of stores to each household. In order to assign triplets of large-surface store choices to each household, we have used the probabilities of selection assigned by the model developed in the first three chapters of this thesis in order to draw the three choices of large-surface store for each household. We begin by drawing each choice of large-surface store independently, using the probabilities assigned by our model. We

have adjusted the draws in order to ensure that when no choice of large-surface store is drawn for one of the choices, no choice is drawn in any subsequent choice.

4.3.3 Imputation with store choices known

In Table 4.15, we suppose that store choice is known for every individual, and we compare the results of imputation using the characteristics of the choices of stores made by each household as auxiliary variables. The control case is the case where we assume that households choose two large-surface stores, and in both these stores, they select products from every single food category. The other cases are different combinations of store type and order of store choice that we have used in order to define imputation classes. We begin by using imputation without using any imputation classes. We follow this by using only the number of stores selected by each household as the auxiliary variable. As we can see, this is one variable that makes a difference in imputation, increasing the proportion of assigned product choices that are correct. All the other variables entered into the model are crossed with this variable. By "doublet", we mean that we have defined an imputation class for every combination of large-surface store type for the first two choices of large-surface store, and by "triplet", we mean the same, only for the first three choices of large-surface stores. By "TripNonord", we mean that we have taken into account every combination of sets of types of large-surface store chosen within the three large-surface stores chosen by each household, only we do not take into account the order of these store types. Thus a household who chose a supermarket in first choice, and a hypermarket for the other two choice would be in the same class as someone who chose a supermarket in its second choice and hypermarkets for the other two choices. Since we observed earlier that probabilities of product selection are much more correlated with whether or not a store is a hard discount store than any other type, we introduced the use of the HD1,

Model	PP	B&P	F&V	M&D	F&S	FF	OP	SSD	SSD(S)	# Imp
Control	0.65	0.86	0.82	0.74	0.75	0.98	0.50	74521	82763	12573
No Vars	0.61	0.63	0.61	0.61	0.60	0.71	0.24	73.2	14101	12573
NumStores	0.66	0.78	0.73	0.69	0.69	0.96	0.40	71.0	13799	12573
doublet	0.68	0.78	0.74	0.70	0.69	0.96	0.40	51.8	13338	12573
doubnonord	0.68	0.78	0.74	0.70	0.69	0.96	0.40	51.8	13338	12573
triplet	0.68	0.79	0.74	0.71	0.70	0.96	0.40	75.0	14359	12573
tripnonord	0.67	0.78	0.74	0.70	0.69	0.96	0.40	62.6	13267	12573
HD1	0.67	0.78	0.74	0.70	0.69	0.96	0.40	98.6	13750	12573
HD2	0.67	0.79	0.74	0.70	0.70	0.96	0.40	73.2	14264	12573
HD1,HD2	0.67	0.79	0.74	0.70	0.69	0.96	0.40	66.8	13971	12573

Table 4.15: Imputation by imputation class using characteristics of known store choices

a dummy variable indicating whether a hard discount store had been chosen in the first choice of store and HD2, which represented a dummy variable indicating whether a hard discount store had been chosen in the second choice of store.

As we can see, the well-allocated statistics for each product individually are actually much higher when we assume households select every single alternative shown in the first row, but this comes at the sacrifice of badly representing the breakdown of households' different distributions of product choices, as is shown by the SSD terms at the right. We find, unsurprisingly, that when we take into account the number of stores visited by each household, we are better able to predict accurately its choices of food category. However, there is no significant improvement in the SSD statistics, which are overall market shares of each combination of food product categories.

Unfortunately, this visible improvement in our model is impossible to achieve in practice, since we wish to apply imputation to households for whom the choices of large-surface stores are unknown. We therefore attempt to exploit our earlier model of store choice in order to gain at least some improvement in our model.

Imputation with store choices unknown

We begin by treating each of the three store choices of each household as if they were independent of the other store choices for the same household, as we do in our model of store choice. We make use of the probabilities of store selection developed in this thesis in order to assign predicted probabilities of selecting each store. With these probabilities, we can draw one store at random from each choice set for each household, which will then be recorded as the predicted store choice for the household in question. These assigned store choices will then be used as the basis of a model in which we predicted the frequency of store visits, and types of products bought.

We note that we use the probabilities assigned in cross-validation as the basis of our store choice draws. When we divide our sample into a receiver and a donor set, the store choice model parameters are estimated on the donor set, and then applied to the receiver set. These choice probabilities would be the basis of the draws of large-surface store choice for each receiver household. The characteristics of these drawn store choices in the receiver sample will be used to find the matching characteristics of the observed store choices of households in the donor sample. In Table 4.16, we do imputation using imputation classes that are determined for receiver households using store choices that are drawn independently.

We can see in Table 4.16 that if our imputation worked better when we made use of the store characteristics in imputation, this is certainly not the case when we use characteristics of assigned store choices. In every measure, our imputation model works worst when we use imputation classes, than when we draw any household at random.

We have difficulty improving our score in cross-validation when we do imputation. We are probably undone by the uncertainty present in the assigning of large-surface stores. This is probably due to the predicted probabilities assigned in our model. Rarely does

Model	B&P	F&V	M&D	F&S	FF	OP	Vect	SSD	SSD(S)	# Imp
Control	0.65	0.86	0.82	0.74	0.75	0.98	0.50	74521	82763	12573
NoVars	0.61	0.63	0.61	0.61	0.60	0.71	0.24	66.4	14710	12573
NumStores	0.59	0.62	0.59	0.59	0.59	0.70	0.21	902.6	17989	12573
doublet	0.59	0.62	0.60	0.59	0.59	0.70	0.21	965.9	18621	12573
doubnonord	0.59	0.62	0.60	0.59	0.59	0.70	0.21	965.9	18621	12573
triplet	0.59	0.61	0.59	0.59	0.59	0.70	0.21	967.1	17684	12560
tripnonord	0.59	0.61	0.59	0.59	0.59	0.70	0.21	943.9	17874	12560
HD1	0.59	0.62	0.60	0.59	0.59	0.70	0.21	924.7	17806	12573
HD2	0.59	0.61	0.60	0.59	0.59	0.70	0.21	954.1	17110	12573
HD1,HD2	0.59	0.62	0.60	0.58	0.59	0.70	0.21	926.3	18615	12573

Table 4.16: Imputation by imputation class using characteristics of drawn store choices

the probability of drawing any one type of store go higher than 0.50, meaning that if our auxiliary variables are wrongly attributed whenever store types are wrongly attributed, the crucial variables in Tables 4.16 will be incorrect for most households, meaning as well that these households will be imputed values from donor households in the wrong imputation class most of the time. Before yielding, however, we must note that some of the inaccuracy in the assigning of large-surface stores to households comes from our false assumption of independence of household choices of stores.

In Tables 4.17 and 4.18, we show in our sample of 12,573 households, how many households are observed selecting each type of store for their first and second choices of stores, and how many choices of each type of store are assigned. We see that the store types of the observed and assigned store choices are roughly equally distributed within the population.

However, we have a problem when we look at pairs of store choices, as in Table 4.19. That there is a large difference between the distribution of store types among these pairs of store choices comes undoubtedly from correlations between the store type

Store Type	Observed	Assigned
OUT	2262	2173
SM	4689	4792
HM	3083	3059
HD	513	524
XM	1808	1811
NO	218	214
Total	12573	12573

Table 4.17: Types of stores of first store choice

Store Type	Observed	Assigned
OUT	2256	2261
SM	2635	2617
HM	1870	1866
HD	792	794
XM	1769	1806
NO	3251	3229

Table 4.18: Types of stores of second store choice

of the first choice of store and the store type of the second choice of store. The most flagrant example of this occurs due to the fact that in the construction of our data, when there was only one choice of store, it was recorded as being the first choice, thus a household could not be recorded as selecting a store in its second choice if its first choice was a nonchoice. However, such pairs of store types could be assigned if we drew our alternatives independently in the first two choices of stores. We also see that there is a positive correlation between households selecting a store with an unknown store type in the first and selecting an unknown store type in the second. Finally, we see that households tend to be more likely to choose different store types in their two choices of stores than choose the same store type.

We have tried several options to deal with this short of rebuilding our store choice model that incorporates the correlations between store choices in the utility expression. We have tried to model the second choice of large-surface store as being dependent on the first choice. We tried modeling the first choice of store, then drawing the first choice, and using the characteristics of the first assigned store, (for example, the type of store), as explanatory variables for the second drawn store choice. We also tried running a different conditional logit model for each store type of the first choice of large surface store. A fourth option we tried was using a clustering procedure to group households by the vector of predicted probabilities of selecting each alternative in the first choice of large-surface store. In each of these subpopulations, we would then create a separate conditional logit model of store choice. Of these four methods, the clustering procedure seemed the best, however, none of these options guaranteed a distribution of types of stores associated with each pair of store choices that would even closely resemble those observed.

The way to account for the correlations between choices of large-surface store types

1st	2nd	Observed	Assigned
OUT	OUT	611	0
OUT	SM	406	1
OUT	HM	307	6
OUT	HD	94	22
OUT	XM	336	49
OUT	NO	508	2095
SM	OUT	699	1091
SM	SM	861	1693
SM	HM	752	721
SM	HD	290	341
SM	XM	735	572
SM	NO	1352	374
HM	OUT	504	769
HM	SM	748	550
HM	HM	474	792
HM	HD	248	222
HM	XM	368	378
HM	NO	741	348
HD	OUT	80	128
HD	SM	152	84
HD	HM	116	82
HD	HD	31	49
HD	XM	86	118
HD	NO	48	63
XM	OUT	362	266
XM	SM	468	280
XM	HM	221	243
XM	HD	129	143
XM	XM	244	652
XM	NO	384	227
NO	OUT	0	7
NO	SM	0	9
NO	HM	0	22
NO	HD	0	17
NO	XM	0	37
NO	NO	218	122

Table 4.19: Comparison of observed and predicted number of households for different combinations of store types in first two choices of large-surface stores.

that we present here is an ad hoc solution that adjusts the predicted probabilities of selection of the three choices of large-surface stores according to the observed distribution of store types. We will only look here at the first two choices of large-surface stores, the methods described here being easily extended to three store choices.

Let j be a choice of store within a given household's set of alternatives J_n, and $F(j)$ be the type of store of store j. P^1_{nj} is the probability of household n choosing store j for its first choice, and P^2_{nj} is the probability of the same household choosing the store for its second choice. Since we consider stores of the same type interchangealbe, what concerns us here will be the probabilities of selecting a store within a store type. Thus, $P^1_{nF(j)}$ and $P^2_{nF(j)}$ will be the probabilities of selecting stores of the same store type as store j in the first and second choices of stores, respectively. $P^{12}_{nF(j)F(k)}$ is the probability of household n choosing the store type of store j as its first choice, and store type $F(k)$ as its second. Naturally,

$$P^{12}_{nF(j)F(k)} = P^1_{nF(j)} P^{2|1}_{nF(k)|F(j)}$$

where $P^{2|1}_{nF(k)|F(j)}$ is the probability of household n selecting store type $F(k)$ as its second store choice provided that it chose $F(j)$ as its first store choice.

If $O^{12}_{F(j)F(k)}$ is the number of households in our sample of households N who are observed choosing store type $F(j)$ in their first choice, and store type $F(k)$ in their second, then we would expect

$$\sum_{n \in N} P^{12}_{nF(j)F(k)} \approx O^{12}_{F(j)F(k)} \forall j, k \in J_n \forall n \in N$$

We have shown above that

$$\sum_{n \in N} P^1_{nF(j)} P^2_{nF(k)} \neq O^{12}_{F(j)F(k)}$$

which can be explained by the fact that

$$P^1_{nF(j)} P^2_{nF(k)} \quad \neq \quad P^{12}_{nF(j)F(k)}$$

or in other words, that

$$P^2_{nF(k)} \quad \neq \quad P^{2|1}_{nF(k)|F(j)}$$

What we hope to find is a simple adjustment of P^2_{nk}, call it $\tilde{P}^{2|1}_{nk|j}$ such that

- $\tilde{P}^{2|1}_{nk|j}$ is between zero and one $\forall$ n and k
- $\sum_{n \in N} P^1_{nF(j)} \tilde{P}^{2|1}_{nk|j} = O^{12}_{F(j)F(k)}$
- $\sum_{n \in N} \tilde{P}^{2|1}_{nk|j} = O^2_{F(k)}$
- $\tilde{P}^{2|1}_{nk|j}$ would depend as much as possible on $P^2_{nF(k)}$

Our solution was to find a constant $\alpha_{F(j)F(k)}$ for each combination of store types for the first choice of store, and the second choice of store, so that

$$\tilde{P}^{2|1}_{nk|j} \quad = \quad \frac{\alpha_{F(j)F(k)} P^2_{nk}}{\sum_{l \in J_n} \alpha_{F(j)F(l)} P^2_{nl}}$$

is equivalent to

$$\tilde{P}^{2|1}_{nk|j} \quad = \quad \frac{e^{\ln \alpha_{F(j)F(k)} + \ln P^2_{nk}}}{\sum_{l \in J_n} e^{\ln \alpha_{F(j)F(l)} + \ln P^2_{nl}}}$$

which is in fact a Conditional Logit model that can be calculated on our data set where we fix the coefficients of the log of the $\ln P_{nk}^2$ at one. For the purpose of estimation, we enter the type of store of the first observed store choice for the household in question, in order to obtain an estimate of the adjustment factors $\alpha_{F(j)F(k)}$ that we will enter into our model. When we assign our store choices, therefore, we can select the first alternative for each store choice for each household, using as sample weights the probabilities already assigned by our original model of store choice. We then base our adjustment factors on the first choice of store assigned, and then draw our second choice of store using as our sample weights the second choice of store assigned multiplied by the adjustment factor.

Unfortunately, we find that this technique allows us to adjust the predicted probabilities so that the sums of the predicted probabilities of selecting any alternative, or group of alternatives over the entire population exactly matches the number who observed this. Unfortunately, so far, we have found that under cross-validation, these adjustment parameters are not stable and we fail to regenerate the market shares of the different store type combinations. This means that the integration of correlations between different choices demands more sophisticated solutions. The reader is advised to look at Turolla (2007 [48]) for an example of how multiple choices of grocery stores can be integrated into a single model.

In a final note, we attempt using Nearest Neighbour Imputation using the probabilities assigned to each store choice assigned by our Conditional Logit models. We use the probabilities calculated for the "3232" model in Section 3. Thus, for each household we have 36 scores, which are the probabilities of selection alternative in each of the three choice sets that were assigned by this model. Since these probabilities are constant for all households sharing the same "IRIS", when we do imputation, we identify the "IRIS" with associated probabilities that are nearest the receiver household in question, and we

Model	B&P	F&V	M&D	F&S	FF	OP	Vect	SSD	SSD(S)	# Imp
Control	0.65	0.86	0.82	0.74	0.75	0.98	0.50	74521	82763	12573
No Vars	0.60	0.59	0.58	0.58	0.57	0.66	0.22	65.3	13904	12573
NN	0.60	0.62	0.61	0.60	0.60	0.71	0.24	995.0	52876	12573

Table 4.20: Results of nearest neighbour imputation compared with imputation through unrestricted random selection of any donor household.

draw a household at random from this "IRIS", and use it for imputation. This procedure is much more time-intensive than the other imputation techniques, since it requires the calculation of an enormous number of distance functions. However, after evaluating the results of imputations we conclude that this technique is not more beneficial than the other techniques we described in this chapter. In Table 4.3.3, we compare Nearest Neighbour imputation with imputation done without any imputation classes.

4.4 Summary

This chapter was devoted to attempting to use imputation techniques in order to assign choices of food product categories that households would purchase in each of their three possible choices of large-surface grocery store. We began by looking at how food product categories were chosen in the Centre Region. Unlike store choices, these decisions were not influenced by the distance of the store in question. We found that choices of food products were correlated, shoppers often selecting either most or all food product categories in one store, or only selecting one or two categories. We also found that the choice of food product category in the first, second, and third choices of large-surface stores were also often correlated. When a household buys products from one category, it often bought from the same category in its other purchases.

In order to assign predictions of food category choices to households, we used deck

imputation. Briefly put, this type of imputation consisted of assigning to an individual the response recorded for another individual taken from another data set. We decided that because the choices of product categories associated with each choice of store were related, that we needed to impute all choices of food products for all three possible choices of large-surface stores from the same individual.

Here, we described several ways in which we could select this donor household, and tried them on our data. We used imputation by imputation class, by classes constructed using model scores, and by nearest neighbour imputation. In each case, we evaluated our imputation techniques by using cross-validation by department on our survey region.

On our data, we began by using imputation only on socio-demographic variables, since the choices of large-surface stores remained unknown for each household. We tried using imputation classes that we constructed by crossing sociodemographic variables that we determined through Logistic regression were the most significant determinants of household behaviour. We also used a clustering procedure on model scores from our Logistic Regressions to create imputation classes. In both cases, none of the auxiliary variables improved the results of our imputation beyond that which was achieved when we would select one individual completely at random from the entire donor set.

We found that we could improve the results of imputation if we used the number of large-surface stores visited by households to determine imputation classes. However, we wished to develop an imputation strategy that could be used in cases where this information was unknown. We attempted imputation based on the characteristics of stores assigned to households through draws following the predicted probabilities of store selection that we developed in this thesis, and nearest neighbour imputation, also based on these probabilities. Neither of these options were viable.

In conclusion, we find that if we use imputation in order to assign choices of large-

surface stores to households, we do not improve our results through the use of auxiliary variables, if we do not know the household's shopping behaviour. However, the model can be improved through the introduction of nothing more than how many stores the household visits. This is something that future econometricians can take into account when they wish to predict this type of choice when they wish to make projections over a population whose behaviour is unknown.

Conclusion

During the course of our thesis, we have developed a model of store choice that we believe can be used as a tool for predicting the shopping flux between geographic regions and large-surface shopping stores that was observed in a survey done by BVA of the Centre Region of France. The model is relatively simple to use, has rapid calculation times, and seems to provide a satisfactory degree of predictive accuracy. If combined with the use of imputation, it can provide forecasts of the number of people buying each category of product within each large-surface store, a level of detail that is very remarkable. Such a model can be combined with other models already developed by BVA of the amount of spending done in general on each type of product to create an idea of exactly how much money an individual in a given location will be expected to spend on a given product. Such a model will be a great benefit to clients involved in grocery retail in the development of their business model, and in their choice of large-surface store emplacement.

We have developed a Conditional Logit model in order to predict large-surface store choice, and an imputation technique for the choice of food product categories based on the choice of large-surfaced store. After investigating many options, we have found that the model that we would recommend begins with a choice set for each individual defined as containing the closest stores to the household's home within each category of large-surface store, along with an "outside option" referring to the choice of all other stores and

a "no store" option referring to the choice of no stores. The number of different stores to include in this choice set will depend upon the computational resources disposed of by BVA, but we recommend having the "outside option" account for no more than about 5 percent of store choices. The predicted choice of store for each household will be predicted using the Conditional Logit model.

If we are dealing with multiple choices of large-surface stores, we model each order of store choice as a decision taken independently of the other store choices. In doing so, we must be able to incorporate the interaction between store choices if we are to use the predicted probabilities of store selection as the basis for drawing alternatives that we wish to assign to each individual. If we wish to forecast, say, the three large-surface stores visited by each household, then we must draw the first choice of store, and then adjust the probabilities of drawing the subsequent alternatives based on the first choice.

We recommend investing in obtaining as much geographic information as possible to serve as explanatory variables in the discrete choice model. When dealing with a data set as large as the one provided by BVA, it is surprising how fine the distinctions represented by the significant parameter estimates of the model are. These can provide insights into shopping behaviour.

When moving beyond predicting store choice, and forecasting food product category choice, we recommend the use of imputation. We did not find auxiliary variables that could be used to improve imputation, but we believe that if a survey company invested in the resources required to find how many stores each individual chose, this would significantly improve the quality of the imputation.

The process of arriving at this model has highlighted several important concerns for us:

- **The primacy of geography over socio-demographic variables in the pre-**

diction of store choice We have found that an effective model of grocery store choice must depend on geographic and spatial factors such as the spatial distribution of large-surface stores, population density, transportation networks, and economic activity. The socio-demographic characteristics of households do not have explanatory value when introduced in such a model.

- **The importance of defining a realistic consideration set of large-surface stores** When predicting the choice of large-surface stores, the definition of the choice set in the choice model is extremely important. We have found that for any given household, the number of large-surface stores for which there is a nonzero probability of selection is enormous. However, attempts to create choice sets broad enough to take into account every event, including the most rare, tend to reduce the accuracy of the model's predictions. It is far better to reduce the number of alternatives in the choice of large-surface stores to only those stores that have a minimum likelihood of selection, even if that means that our model will not assign probabilities to rare events, than to attempt to assign probabilities to the selection of every store, through the use of sampled alternatives, or through the use of a simpler gravitational model. As a rule of thumb, we ought to be including in a household's choice set only those alternatives that a typical individual will choose, that is, those stores in the household's immediate neighbourhood. Adjusting our choice set to incorporate all possible choices of stores, including those made by households in exceptional circumstances (who have recently moved, but retain a job in their old home hundreds of km away, who buy their groceries where they regularly visit their relatives, etc.) is counterproductive.

- **How to evaluate models of large-surface store choice** The evaluation of our model will depend upon the use to which the model is put. In our case, we wish

to use our discrete choice model in order to predict the number of clients visiting each large-surface store. We therefore created a metric, that we called the "WD" statistic that is based on the differenced between the forecast and observed number of clients from each geographic region visiting each large-surface store. Such results could allow us to use a criteria that is not necessarily represented by more standard measures of model quality, such a goodness of fit measures.

- **Forecasting by imputation: only as good as our large-surface store choice model allows it to be.** We have found that forecasting food product categories by imputation only improves over randomly assigning food product categories when we use as auxiliary variables the number of stores and store types selected by each household. Since these remain unknown, the reliability of our method will depend upon the ability of our store-choice model to predict the stores chosen by the household.

There remain many open-ended questions and possibilities for improvement at the end of our thesis.

Possibilities of model structure improvement One of the advantages of our model is that it is fairly rapid to calculate with respect to alternative models. We believe that the pursuit of more general model forms could compromise this efficiency in terms of computational time. As well, the structural simplicity of our model is compensated by the wealth of information contained wihtin the explanatory variables used. We believe that in general, if we wish to take more detailed considerations into account, it is better to do this through a more detailed data base, than by rendering our model more complex. However, there are a few obvious ways in which our model can be improved structurally:

- **Random taste variation.** We have stressed the point that explanatory variables

representing differences in the characteristics of different households do not have explanatory power in our model. However, that does not mean that individual taste variation is not an important factor in household decision making. What could be the case is that this taste variation is not captured by our socio-demographic variables. What can be done to account for this is to suppose that our model parameters vary by individual, following a known random distribution. This is what is commonly referred to as a Mixed Logit model, and is frequently found in the literature. An example of a recent paper that discusses this type of model is Walker, Ben-Akiva, Bolduc, (2007, [49]) that discusses Normal Error Component Logit Mixture (NECLM) models. The problem with this model type is that it takes much longer to calculate than a Logit model with fixed parameters. Our model already contains a very large number of parameters, and so the calculation of such a model will take a very long time.

- **Spatial auto-correlation of error terms.** In order to take into account the effect of competition or market cannibalization, we have considered structuring our model so as to introduce a correlation between utilities that depends upon the euclidean distance that separates them. A discrete choice model that explicitly takes into account spatial interaction has been introduced by Lesage (2000, [26]), which is further developed in a paper by Smith and Lesage (2004, [43]). We have not used such models, but we have entered many explanatory variables that we believe helps account for market competition, and the effect of proximity between chosen stores.

- **Multiple discrete choice models.** A last possibility is the prediction of all three choices of large-surface stores in a single step. This has not yet been done. What must be determined is whether such a model will greatly increase computation

time.

- **Better data will be available.** During our thesis, we have repeatedly asserted that we were pleased to have completely geocoded household and large-surface store locations, which allowed a level of specification that was not possible in earlier years. However, survey institutes like BVA are steadily improving the amount and quality of data available. In the courses of our studies, BVA produced calculations of estimated travel times by motor vehicle between every single IRIS within France. This arrived too late for it to be implemented in our thesis, but it would provide an improvement to our use of the euclidean distance as a proxy of the time required for each household to visit each large-surface store. Initial uses of this new data did not reveal a substantial difference in our parameter estimates, and we find that the ranks of store distances by vehicular travel times often correspond with the ranks of the euclidean distances of stores, but using this more accurate measure of distance could allow our model of store choice to be more readily transferable to other regions in which the relationship between euclidean distance and travel time is different (for example, in mountainous regions, where travel times tend to be slower). This is a hypothesis that needs to be tested. Other areas in which we new relevant data could become available is with respect to grocery shopping outside large-surface stores. Households were questioned on their grocery purchases in traditional commerce, and marketplaces, but their shopping locations had not yet been geolocalized and thus we could not study the spatial dimension of this type of shopping. As well, we have used many variables as proxies for the possible effects of home-work trajectories. BVA has invested a great deal in order to obtain data, and analysis of the home-work trajectories of households. Future models of shopping behaviour will undoubtedly include such information. A last and most

important point in which more relevant data can be obtained is with respect to prices and household spending. Not present in our work here, such information would prove very useful if it can be obtained for any future studies of shopping behaviour.

The transferability of our model to other geographic regions

A last, and vital question is the ability of our model to make accurate predictions on another French region. From our cross-validation, we can conclude that our model would work well if applied to another region, if the transportation and urban structure were similar. However, the landlocked Centre Region of France is mostly covered by plains, and has a roughly homogeneous distribution of medium-sized cities. Other Regions of France are quite different from the Centre Region geographically, and contain features that are not found in Centre and could not be taken into account. Some regions, such as Auvergne, and Rhone-Alpes are quite mountainous, having road networks that are slower. The Parisian Region contains the 9 million people of Paris, the second largest urban agglomeration in Europe. The Provence-Alpes-Côtes-d'Azur Region also has a very distinctive geographic structure, containing a very large population occupying a long, heavily urbanized narrow strip along the southeast coast of France. Due to the primacy of geographic structure in our model of store choice, this variation between regions will have to be addressed if we wish to use our model to predict store choices everywhere in France. We have begun to look at how our model behaves in the neighbouring region of Auvergne, and we did find that the model performed less well. Further studies will need to be done in order to see how our model quality varies by region in France. We recommend that if we are to create a model of store choice that applies to all of France, that some representative samples will have to be obtained of different parts of France so that different geographic realities be taken into account. Rather than taking five

departments from the same region that are very similar in character, we may do well to select several different departments selected from different parts of France.

If we need to worry about how much a model's accuracy varies with geography, the variability of the model's accuracy over time is just as much of a concern. As we noted in our introduction, the grocery store market in France is undergoing an important transformation, with the market shares of hard discount stores changing significantly from year to year and large numbers of new large-surface stores constantly being installed. As well, we must also note the significant changes in household behaviour, and household displacement patterns, which are so important in determining the success of our model. Another effect is the introduction of hard discount stores. These stores are rapidly being implanted everywhere in France, and they are increasing their market share. We have already seen that hard discount stores have a very different attraction on consumers than other types of stores. If more consumers start going to hard discount store, this will mark a shift in consumer behaviour. Because our data comes entirely from a single survey of shopping behaviour undertaken in 2004, we will need to look at how our model behaves with new data taken over time. We believe that due to changes in French society, the type of model that was developed in this thesis will constantly need to be updated with new data.

As time goes by, we expect that French shopping behaviour will become more and more difficult to predict, due to the increasing complexity of household displacements, but that does not mean that behaviour forecasts become less worthwhile. We believe that because of this greater unpredictability the importance of consumption forecasting and that the interest that players in the grocery market will only increase. However, just as GPS maps in cars rely on the gathering of great quantities of reliable and up-to-date data, we believe that the type of forecasts that we developed here will also require

investment in large quantities of constantly updated and accurate data. We hope that in the future, the type of model here developed will be found to be as useful and ubiquitous as the new GPS digital location devices found in cars.

Sébastien Markley

Toulouse, France

Glossary 1: Abbreviated terms in figures and tables

18	The Cher Department
2121	A reference to a choice set or any model referring to a choice set containing the closest two supermarkets, the closest small hypermarket, the closest two hard discount scores, and the closest large hypermarket.
28	The Eure-et-Loir Department
3232	A reference to a choice set or any model referring to a choice set containing the closest three supermarkets, the closest two small hypermarkets, the closest three hard discount scores, and the closest two large hypermarkets.
36	The Indre Department
37	The Indre-et-Loir Department
41	The Loire-et-Cher Department
4372	A reference to a choice set or any model referring to a choice set containing the closest four supermarkets, the closest three small hypermarkets, the closest seven hard discount scores, and the closest two large hypermarkets.
7694	A reference to a choice set or any model referring to a choice set containing the closest seven supermarkets, the closest six small hypermarkets, the closest nine hard discount scores, and the closest four large hypermarkets.
BrierZ	Brier Score of averages of predicted probabilities over geographic zones
Calib	Calibration, a component of the Brier Score
CalibAbs	Calibration measured using absolute difference between observed and predicted shopping flows, instead of the square of the difference.
CalibNO	Calibration, calculated without reference to choices of no stores
Cont	Continuous variable

ContLin	The control case for model estimation that is used to serve as a baseline for the evaluation of the gravitational model parameter estimates. Here, it is assumed that the probability of any household selecting any store is equal to the percent of households in the entire region who select that same store. In other words, the probability of selecting any store is constant for all individuals.
CovByZ	Within-zone covariance of predicted probabilities and observed response
Ctrl	Control case
Dich	Dichotomous, or binary variable
doublet	A pair composed of the type of stores of the first and the second choices of large-surface stores made by a household.
doubnonord	The types of stores of the first and second choices of large-surface stores made by a household, in which the order in which the store types comes is not taken into account.
Grav	Gravitational model of store choice
HD	Hard Discount Store
HD1,HD2,HD3	Dummy variables indicating that a hard discount store was selected in either the first, second, or third choice of large-surface stores, respectively.
HM	Small Hypermarket
# Imp	Number of households whose values were imputed
Ind	Individual
INSEE	*Institut National de la Statistique et des Etudes Economiques*, the French national bureau of statistics
Iri/IRIS	*Ilots regroupés pour l'information statistique*, a geographic subdivision of France for the purpose of data collection
Low Var	A measure of the variability of the Logit model parameters calculated in cross-validation.
NO	Corresponds to the choice of no store in our models of store choice.
N05Lin	The gravitational model of store choice in which parameters were calculated using a regression of empirical probabilities of selecting stores calculated over sets of household-store-choice pairs that had on average 5 members.
N10Lin	The gravitational model of store choice in which parameters were calculated using a regression of empirical probabilities of selecting stores calculated over sets of household-store-choice pairs that had on average 10 members.
N20Lin	The gravitational model of store choice in which parameters were calculated using a regression of empirical probabilities of selecting stores calculated over sets of household-store-choice pairs that had on average 20 members.
N50Lin	The gravitational model of store choice in which parameters were calculated using a regression of empirical probabilities of selecting stores calculated over sets of household-store-choice pairs that had on average 50 members.
N100Lin	The gravitational model of store choice in which parameters were calculated using a regression of empirical probabilities of selecting stores calculated over sets of household-store-choice pairs that had on average 100 members.

New Sgn	The number of times that a parameter estimate calculated in cross-validation is significant and of a different sign than the estimate of the same parameter in the entire Centre Region
New Sig	The number of parameter estimates that are not significant when estimated over the entire Centre Region, but are significant in at least one of cross-validation steps.
NG12	A reference to a choice set constructed by doing 11 draws in SRS random sampling of alternatives.
NG18	A reference to a choice set constructed by doing 18 draws in SRS random sampling of alternatives.
No Vars	No variables used in order to constrain the draws of donor households for the purpose of deck imputation.
Not 18	The subset of our sample of households that resides in any department of the Centre Region but the Cher Department.
Not 28	The subset of our sample of households that resides in any department of the Centre Region but the Eure-et-Loir Department.
Not 36	The subset of our sample of households that resides in any department of the Centre Region but the Indre Department.
Not 37	The subset of our sample of households that resides in any department of the Centre Region but the Indre-et-Loir Department.
Not 41	The subset of our sample of households that resides in any department of the Centre Region but the Loire-et-Cher Department.
NumStores	The number of large-surface stores visited by a household.
OUT	Corresponds to the outside option in our models of store choice.
R18	A reference to a choice set constructed by doing 17 draws of PPS random sampling of alternatives.
R31	A reference to a choice set constructed by doing 30 draws of PPS random sampling of alternatives.
Resol	Resolution, a component of the Brier Score
Sec	Survey sector
SM	Supermarket
SSD	The weighted sum of squares of the difference between the observed number of households selecting each combination of food product categories and the number assigned through imputation
SSD(S)	The weighted sums of squares of the difference between the observed number of households in each survey sector selecting each combination of food product categories and the number assigned through imputation
triplet	An ordered set of composed of the type of stores of the first and the second choices of large-surface stores made by a household.
tripnonord	The types of stores of the first and second choices of large-surface stores made by a household, in which the order in which the store types comes is not taken into account.
UV	Uncontrolled Variation, a component of the Brier Score
VarByZ	Within-zone variance of predicted probabilities

Glossary 2: Sociodemographic variables

AccVilFCat	5 quantiles of the variable AccVilF.
actcat	Identifies whether the person of reference for the household or the conjoint of the person of reference is active.
achicat	avgnachi recoded as a categorical variable.
age	Age of person of reference for the household.
agecat	the age variable recoded in categorical form.
agemoy	Average age of individuals in household.
anlgcat	avgnlogn recoded as a categorical variable.
avgnlogn	Average number of dwellings per residential housing unit in the IRIS of the household in question.
avgnachi	Average of the indexes of the periods of construction for the residential housing units in the IRIS of the domicile of the household in question.
carcat	Whether there is at least one person declared adult in the household who drives a car to work.
carcat2	Whether the person of reference drives a car to work.
cle1	Identifies the household in the sample.
cplcat	Whether the person of reference of the household is in a couple or not.
cspcat	Whether the person of reference is in a "higher" or "lower" socioprofessional category, or is inactive.
dciris	Identifies an IRIS and the department and commune in which the observation refers.
disij	The distance between the centroid of the IRIS of residence of the household in question (the population centre of gravity of the corresponding commune in the case of rural IRIS) and the LS-store.
discat	disij recoded as a categorical variable.

distrnkabs	The rank of the distance of the LS-store from the household in question within a given choice set. For example, if an LS-store is the third closest store to a household within its choice set, distrnkabs has value 3. Ranks go from 1 to 30 but distrnkabs takes value 31 if it is the rank of a store a household chooses, but the distance is over 50 km (and therefore it is not among the stores automatically assigned to the choice set of the household).
drkbytype	The rank of the distance of the LS-store within its category within the choice set. For example if a store is the fourth closest supermarket within a given choice set to the household, drkbytype takes value 4. Ranks go from 1 to 10 possible stores for each category, but takes value 11 if the household chooses a store within the category that is over 50 km away (and therefore is not among the stores automatically assigned to the choice set of the household).
Enfcat	Presence of children in the household.
Enseigne	The name of the LS-store chain.
Ens_id	The numeric code assigned to each value of Enseigne.
exppneetr	The exponential of PNeEtr.
FV	Type of LS-store (1=OUT,4=SM,5=HM,6=HD,7=XM,9=NO)
Hsizecat	Number of people in the houseold.
inactcat	If there is at least one person declared an adult in the household who is inactive.
exppnatetr	The exponential of pnatetr.
Lieuchx	Variable indicating a household's choice of LS-store. It is one if the household in question chose the store in question, and zero otherwise.
lnavgnlogn	The log of avgnlogn.
lnavgnachi	The log of avgnachi.
lndis	The log of disij, zero if unknown.
lndisFV4	equal to lndis when FV is 4 and 0 otherwise
lndisFV5	equal to lndis when FV is 5 and 0 otherwise
lndisFV6	equal to lndis when FV is 6 and 0 otherwise
lnsurf	The log of surface, zero, if unknown.
lnsurfFV4	equal to lnsurf when FV is 4 and 0 otherwise
lnsurfFV5	equal to lnsurf when FV is 5 and 0 otherwise
lnsurfFV6	equal to lnsurf when FV is 6 and 0 otherwise
lnuc	The log of UC.
lntu99n	The log of tu99n.
piedcat2	Whether the person of reference walks to work.
MedUC	Average revenue per unit of consumption (a French unit of measure of family size) per IRIS.
medUCCat	medUC recoded as a categorical variable.
NonAl	Number of facilities justifying a visit to commune for reasons other than grocery shopping(existence of gas station, existence of commercial district, etc.).
piedcat	Whether there is at least one person declared an adult in the household who goes to work by foot.
PnatEtrCat	PnatEtr recoded as a categorical variable.

[t]

PNatEtr	Percent of individuals in an IRIS with a non-French Nationality.
PNeEtr	Percent of individuals in an IRIS born outside France.
PNeEtrCat	PneEtrCat recoded as a categorical variable.
pol99n	the polarity of the commune in which the household lives.
PropAP82Cat	5 quantiles of the percent of households living in homes built from 1982 onwards.
propcat	Whether the household in question owns or rents its domicile.
PropLocCat	5 quantiles of of the proportion households in commune renting their residence.
PropTertCnt	5 quantiles of the proportion of the population involved in the tertiary sector.
quotacat	Category of household defined by whether the household is headed by a couple or not, and whether the head of the household is active. Used to define quotas for the creation of the sample.
redressement	Sample weights of households in survey.
rescat	Type of residence: appartment or individual residence.
revcat	Category of revenue for household.
stylchx	Identifies whether a choice of store made by a household is the first, second, or third closest to the household in question among stores within its category.
stylchx2	Variant of stylchx: Identifies whether the household's choice of store is the closest store within its category or not.
stylchx3	Identifies whether the store chosen by the household is the closest store, the second closest, or third closest store to the household.
surface	The surface area of the LS-store accessible to customers in square meters.
surfcat	Surface recoded as a categorical variable.
transcat	Type of transportation used by person of reference in order to go to work.
transcat2	transcat recoded to aggregate modalities.
tu99n	Population category of commune in which household lives.
tu99n2	tu99n with cities between 20,000 and 100,000 collapsed into one category.
UC	Number of units of consumption in the household.
UCCat	UC recoded as a categorical variable.
VC99n	Whether the commune in which the household lives is a city centre, or a suburb.
vehcat	Whether the household owns a car or truck or not.

Glossary 3: Modalities of sociodemographic variables

actcat	1	Neither person of reference nor conjoint are active.
	2	Person of reference or conjoint is active.
achicat	0	avgnachi missing.
	1	avgnachi less than 2.3.
	2	avgnachi greater than 2.3 and less than 3.
	3	avgnachi greater than 2.3 and less than 3.4.
	4	avgnachi greater than 2.3 and less than 3.9.
	5	avgnachi greater than 2.3 and less than 4.6.
	6	avgnachi greater than 4.6.
agecat	0	Blank
	1	15 years old or less, or invalid entry.
	2	16-35 years of age.
	3	36-45 years of age.
	4	16-65 years of age.
	5	65 or more years of age.
anlgcat	0	avgnlogn missing.
	1	avgnlogn less than 1.01.
	2	avgnlogn greater than 1.01 and less than 2.
	3	avgnlogn greater than 2 and less than 5.
	4	avgnlogn greater than 5 and less than 10.
	5	avgnlogn greater than 10.
carcat	1	No adult in household drives a car to work.
	2	At least one adult in household drives a car to work.
carcat2	0	Person of reference does not drive a car to work.
	1	Person of reference drives a car to work.

cplcat	1	Household headed by a single person.
	2	Household headed by a couple.
cspcat	0	Unknown.
	1	Inactive.
	2	"Lower" socioprofessional category.
	3	"Higher" socioprofessional category.
discat	0	unknown surface area.
	1	From 0 to 500 square meters.
	2	From 500 to 749 square meters.
	3	From 750 to 1,499 square meters.
	4	From 1,500 to 2,999 square meters.
	5	From 3,000 to 6,999 square meters.
	6	From 7,000 to 14,999 square meters.
	7	From 15,000 to 24,999 square meters.
	8	From 25,000 to 49,999 square meters.
ens_id	1	8 A HUIT
	2	ALDI
	3	ATAC
	4	AUCHAN
	5	Aucun
	6	CARREFOUR
	7	CASINO
	8	CENTRE LECLERC
	9	CHAMPION
	10	COCCINELLE
	11	COMOD
	12	CONTINENT
	13	COOP
	14	CORA
	15	DIAGONAL
	16	ECO SERVICE
	17	ECOMARCHE
	18	ED
	19	Erreur
	20	G 20
	21	GALERIES LAFAY
	22	HYPER U
	23	INTERMARCHE
	24	Indéfini
	25	Indépendant
	26	LEADER PRICE
	27	LIDL
	28	MARCHE PLUS
	29	MARCHE U

ens_id	30	MAXIMARCHE
	31	MONOPRIX
	32	MUTANT (LE)
	33	NETTO
	34	NORMA
	35	SHOPI
	36	SITIS
	37	STOCK
	38	SUPER U
Enfcat	0	Nonresponse.
	1	No children within household.
	2	Children within household.
FV	4	Supermarket.
	5	Hypermarket.
	6	Hard discount.
Hsizecat	0	Nonresponse.
	1	1 person in household.
	2	2 people in household.
	3	3 or more people in household.
medUCCat	0	meduc missing.
	1	meduc less than 11,000.
	2	meduc greater than 11,000 and less than 12,800.
	3	meduc greater than 12,800 and less than 14,000.
	4	meduc greater than 14,000 and less than 15,300.
	5	meduc greater than 15,300 and less than 17,400.
	6	meduc greater than 17,400.
piedcat	1	No one in the household goes to work on foot.
	2	At least one person in the household goes to work on foot.
piedcat2	0	Person of reference does not go to work on foot.
	1	Person of reference goes to work on foot.
PnatEtrCat	0	PnatEtr missing.
	1	PnatEtr equals 0.
	2	PnatEtr greater than 0 and less than 0.02.
	3	PnatEtr greater than 0.02 and less than 0.05.
	4	PnatEtr greater than 0.05 and less than 0.08.
	5	PnatEtr greater than 0.08.
PNeEtrCat	0	PnatEtr missing.
	1	PnatEtr equals 0.
	2	PnatEtr greater than 0 and less than 0.05.
	3	PnatEtr greater than 0.05 and less than 0.08.
	4	PnatEtr greater than 0.08 and less than 0.12.
	5	PnatEtr greater than 0.12.

pol99n	0	Missing.
	1	Urban pole.
	2	Monopolarized commune.
	3	Multipolarized commune.
	4	Rural commune.
propcat	1	Proprietor of residence.
	2	Household rents its residence.
quotacat	0	Missing.
	1	Household headed by single, active person.
	2	Household headed by single, inactive person.
	3	Household headed by couple in which both are active.
	4	Household headed by couple in which only one is active.
	5	Household headed by couple in which both are inactive.
rescat	1	Household lives in a single-dwelling housing unit.
	2	Household lives in an appartment in a multi-dwelling housing unit.
revcat	1	Total household revenue less than 750 euros per month.
	2	Total household revenue from 751 to 1200 euros per month.
	3	Total household revenue from 1201 to 1500 euros per month.
	4	Total household revenue from 1501 to 2300 euros per month.
	5	Total household revenue from 2301 to 3000euros per month.
	6	Total household revenue greater than 3000 euros per month.
	7	Missing.
stylchx	1	Closest large-surface store to household in choice set and is a supermarket.
	2	Closest large-surface store to household in choice set and is a hypermarket.
	3	Closest large-surface store to household in choice set and is a hard discount.
	4	Closest supermarket to household in choice set but not closest large-surface store.
	5	Closest hypermarket to household in choice set but not closest large-surface store.
	6	Supermarket that is not closest supermarket to household in choice set
	7	Hypermarket that is not closest hypermarket to household in choice set
	8	Hard discount that is not closest hard discount to household in choice set
	9	Large-surface store whose type is unknown.
	11	Unknown distance (usually in cases where household purchases store more than 50 km away.)
	12	Household does not shop in large-surface store.

Stylchx2	1	Closest supermarket to household in choice set.
	2	Supermarket that is not closest to household in choice set.
	3	Closest hypermarket to household in choice set.
	4	Hypermarket that is not closest to household in choice set.
	5	Hard discount store.
Stylchx3	1	Closest large-surface store to household in choice set.
	2	Second closest large-surface store to household in choice set.
	3	Third closest large-surface store to household in choice set.
	4	Large-surface store with unknown rank of distance to household in choice set.
	5	Unknown rank of large-surface store.
Surfcat	0	unknown surface area.
	1	From 300 to 400 square meters.
	2	From 401 to 650 square meters.
	3	From 651 to 800 square meters.
	4	From 801 to 1050 square meters.
	5	From 1051 to 1350 square meters.
	6	From 1351 to 1700 square meters.
	7	From 1701 to 2500 square meters.
	8	From 2501 to 3500 square meters.
	9	From 3501 to 7500 square meters.
	10	From 7501 or more square meters.
Transcat	0	Missing
	1	Commutes to work on foot.
	2	Commutes to work by bus.
	3	Commutes to work by metro.
	4	Commutes to work by scooter.
	5	Commutes to work by tramway.
	6	Commutes to work by car.
	7	Commutes to work by bicycle.
	8	Commutes to work by other forms of transportation.
Transcat2	0	Missing
	1	Commutes to work on foot.
	2	Commutes to work by public transportation (bus, metro, or tramway).
	3	Commutes to work by two-wheeled vehicle (bicycle, or scooter).
	4	Commutes to work by car.

tu99n	0	Rural commune
	1	Less than 5000 inhabitants.
	2	5000 to 10,000 inhabitants.
	3	10,000 to 20,000 inhabitants.
	4	20,000 to 50,000 inhabitants.
	5	50,000 to 100,000 inhabitants.
	6	100,000 to 200,000 inhabitants.
	7	200,000 to 1 million inhabitants.
	8	Parisian urban unit.
tu99n2	0	Rural commune
	1	Less than 5000 inhabitants.
	2	5000 to 10,000 inhabitants.
	3	10,000 to 20,000 inhabitants.
	4	20,000 to 100,000 inhabitants.
	5	100,000 to 200,000 inhabitants.
	6	200,000 to 1 million inhabitants.
	7	Parisian urban unit.
UCCat	0	UC missing.
	1	UC less than or equal to 1.
	2	UC less than or equal to 1.5 and greater than 1.
	3	UC less than or equal to 2 and greater than 1.5.
	4	UC less than or equal to 2.5 and greater than 2.
	5	UC greater than 2.5.
VC99n	0	Missing
	1	Household without a car.
	2	Household with a car.

Glossary 4: explanatory variables for conditional logit model

Variable	Description	Type
AccAutoR	Time in minutes to access closest autoroute	Cont
AccAutoRZ	AccAutoR is zero or missing	Dich
AccComF	Time in hours to go from domicile to commune of store in question if it is the most frequently visited commune by those living in commune of household	Cont
AccComFZ	AccComF is zero or missing	Dich
AccVilF	Time in hours to go from domicile to commune in question if it is the most frequently visited commune by those living in commune of household and it has more than 10,000 residents	Cont
AccVilFZ	AccVilF is zero or missing	Dich
Denspopu	Population density of commune of large-surface store	Cont
dis	Euclidean distance of store from home in km	Cont
dissq	Square of dis	Cont
FavCom	Commune of store is the commune most-visited by population living in commune of household's home	Dich
FavCom	Store's commune is the most visited by those in commune of household's domicile	Dich
FavVil	Store's commune is the most visited by those in commune of household's domicile and it has more than 10,000 residents	Dich
GSpol99le1	Commune of large-surface store classed as urban pole	Dich
GSpol99le2	Commune of large-surface store classed as urban pole or monopolarized	Dich
GSpol99le3	Commune of large-surface store classed as urban pole, monopolarized, or multipolarized	Dich
GStu299le0	Commune of large-surface store classed as rural	Dich
GStu299le1	Population of large-surface store nonrural with less than 10K inhabitants	Dich

Variable	Description	Type
GStu299le2	Population of large-surface store nonrural with less than 50K inhabitants	Dich
GStu299le3	Population of large-surface store nonrural with less than 100K inhabitants	Dich
gsVC99_1	Commune of large-surface store classed as city centre	Dich
HD	Hard discount store	Dich
HDRankGE2	Hard discount store with rank of distance >= 2	Dich
HDRankGE3	Hard discount with rank of distance >= 3	Dich
HM	Small hypermarket	Dich
HMRankGE2	Small hypermarket with rank of distance >= 2	Dich
Nostore	No store	Dich
outside	Outside option chosen ("other stores")	Dich
OutWHDNumGE3	Outside option for choice set with 3 or more alternatives representing hard discount stores	Dich
...	...	...
OutWHDNumGE12	Outside option for choice set with 12 or more alternatives representing hard discount stores	Dich
OutWHMNumGE3	Outside option for choice set with 3 or more alternatives representing small hypermarkets	Dich
...	...	...
OutWHMNumGE12	Outside option for choice set with 12 or more alternatives representing small hypermarkets	Dich
OutWSMNumGE3	Outside option for choice set with 3 or more alternatives representing supermarkets	Dich
...	...	...
OutWSMNumGE12	Outside option for choice set with 12 or more alternatives representing supermarkets	Dich
OutWXMNumGE3	Outside option for choice set with 3 or more alternatives representing large hypermarkets	Dich
...	...	...
OutWXMNumGE12	Outside option for choice set with 12 or more alternatives representing large hypermarkets	Dich
Samecit	Large-surface store is in same commune as household's residence	Dich
Samedep	Large-surface store is in same department as household's residence	Dich
SameUU	Large-surface store is in same commune as household's residence	Dich
SM	Supermarket	Dich
SMRankGE2	Supermarket with rank of distance >= 2	Dich
SMRankGE3	Supermarket with rank of distance >= 3	Dich
surf	Surface area of supermarket in thousands of m2	Cont
surfsq	Square of surf	Cont
TR2ROU	Percent of population in household's home commune commuting to commune of store in question by a two-wheeled vehicle	Cont
TRCOM	Percent of population in household's home commune commuting to commune of store in question by public transportation	Cont
XM	Large hypermarket	Dich
XMRankGE2	Large hypermarket with rank of distance >= 2	Dich

Appendix 1: Estimation of the Conditional Logit Model

If we are to assign probabilities of selection of alternatives, in which

$$P_{nj} = \frac{e^{\beta X_{nj}}}{\sum_{k \in C_n} e^{\beta X_{nk}}}$$

then we need to determine the values of β that we will use in the expression above. We do this by selecting the parameters of our model that maximize the model likelihood of the observed selections of stores made by the households in our population.

Thus, our estimated model parameters $\hat{\beta}$ will be calculated with the following equation

$$\hat{\beta} = \arg\max_{\beta} P(z_n, \ \forall n \in N|\beta)$$

Where z_n denotes the event that household n selects the alternative it was observed to select. $P(z_n, \ \forall n \in N|\beta)$ is the likelihood function. Since each household's choice is independent, this equation becomes:

$$\hat{\beta} = \arg\max_{\beta} \prod_{n \in N} \prod_{j \in C_n} P_{nj}(\beta)^{z_{nj}}$$

where C_n is the choice set of each individual n and z_{nj} is one when household n selects j and zero, otherwise, and $P_{nj}(\beta)$ is a function that calculates the probability that household n will select store j given a set of parameters β. Maximizing the likelihood is equivalent to maximizing the log of the average likelihood for each individual in the sample, so we shall be using the function LL for our estimation.

$$\begin{aligned} \mathrm{LL}(\beta, N) &= \frac{\ln\left(\prod_{n\in N}\prod_{j\in C_n} P_{nj}(\beta)^{z_{nj}}\right)}{N} \\ &= \sum_{n\in N}\sum_{j\in C_n}\frac{z_{nj}\ln P_{nj}(\beta)}{N} \end{aligned}$$

so that

$$\hat{\beta} = \arg\max_{\beta} \mathrm{LL}(\beta, N)$$

The reason we use this estimation procedure (as opposed to a least squares estimation) is that a classical result states that if a set of individuals (or households) N, drawn exogeneously from the population behaves in conformity with the assumptions underlying a conditional logit model, then

$$\sqrt{N}\left(\hat{\beta}-\beta\right) \xrightarrow{d} \mathcal{N}\left(0, (-H)^{-1}\right)$$

where we take β to be the true model parameters, and $\hat{\beta}$ to be the maximum likelihood estimate of β, maximizing $\mathrm{LL}(N, \beta)$. H is the Hessian matrix of the second-order partial derivatives of the average likelihood of any individual selecting the store it has been observed to select.

Provided our likelihood function is globally concave, we can find the vector β that provides maximum values for the likelihood function by calculating the gradient of this

function with respect to β, and setting this equal to zero using Newton-Raphson estimation.

Train (2003 [46]) shows a way in which this estimation is analogous to the OLS regression. We can see this by observing the gradient of the likelihood:

$$\begin{aligned}
\nabla_\beta \hat{P}_{nj}(\beta) &= \nabla_\beta \left(\frac{e^{\beta^T X_{nj}}}{\sum_{k \in C_n} e^{\beta^T X_{nk}}} \right) \\
&= \left(\frac{X_{nj} e^{\beta^T X_{nj}} \sum_{k \in C_n} e^{\beta^T X_{nk}}}{\left(\sum_{k \in C_n} e^{\beta^T X_{nk}} \right)^2} - \frac{\sum_{k \in C_n} X_{nk} e^{\beta^T X_{nj}} e^{\beta^T X_{nk}}}{\left(\sum_{k \in C_n} e^{\beta^T X_{nk}} \right)^2} \right) \\
&= \left(X_{nj} \hat{P}_{nj}(\beta) - \sum_{k \in C_n} X_{nk} \hat{P}_{nj}(\beta) \hat{P}_{nk}(\beta) \right) \\
&= \left(X_{nj} - \sum_{k \in C_n} X_{nk} \hat{P}_{nk}(\beta) \right) \hat{P}_{nj}(\beta)
\end{aligned}$$

and so

$$\begin{aligned}
\nabla_\beta \mathrm{LL}(\beta, N) &= \nabla_\beta \left(\sum_{n \in N} \sum_{j \in C_n} \frac{1}{N} z_{nj} \ln \hat{P}_{nj}(\beta) \right) \\
&= \sum_{n \in N} \sum_{j \in C_n} \frac{1}{N} \frac{z_{nj}}{\hat{P}_{nj}(\beta)} \nabla_\beta \hat{P}_{nj}(\beta) \\
&= \sum_{n \in N} \sum_{j \in C_n} \frac{1}{N} \frac{z_{nj}}{\hat{P}_{nj}(\beta)} \left(X_{nj} - \sum_{k \in C_n} X_{nk} \hat{P}_{nj}(\beta) \right) \hat{P}_{nj}(\beta) \\
&= \sum_{n \in N} \sum_{j \in C_n} \frac{1}{N} (z_{nj} - \hat{P}_{nj}(\beta)) X_{nj} \qquad (4.1)
\end{aligned}$$

Since we suppose that the draw of households is exogeneous, this implies that our model residuals $r_{nj} = z_{nj} - P_{nj}$ are independent of the explanatory variables, and thus, the covariance between z_{nj} and r_{nj} is zero. From the derivation above, we see that the maximum likelihood estimates of β are in fact the values for which the estimated

covariance between residuals and explanatory variables is zero. Least squares estimates in linear regression also contain this property of setting sample covariance between the residual and the explanatory variables to zero. This shows the close similarity between ordinary linear regression and maximum likelihood estimates of the conditional logit model.

Appendix 2: Computation of the Conditional Logit Model Estimates

In order to find the values of β that maximize the maximum likelihood function, we find values of β that set the derivative of the log likelihood function with respect to β to zero. In order to do this, we use the Newton-Raphson method. In this method, if we have a multinomial function f, and we wish to find a vector x at which it attains zero, we begin with an arbitrary starting value x_0. At each step of our iteration, we begin with the vector produced in the previous step, x_n, and we calculate x_{n+1} by taking the zero of a linear approximation of f at x_n. Suppose that

$$J_f(x_n)(x_{n+1} - x_n) \approx f(x_{n+1}) - f(x_n)$$

where J_f is the Jacobian matrix of first order derivatives of f at x_n. Setting $f(x_{n+1})$ to zero, we obtain the equation

$$x_{n+1} = x_n - [J_f(x_n)]^{-1} f(x_n)$$

that we use to find x_{n+1} Applying this to find the zero of the gradient of the log-likelihood function, noting that the Jacobian of a gradient is a Hessian matrix, we begin with an

arbitrary value of β, call it β_0 and we use the following equation to make successive iterations

$$\beta_{n+1} = \beta_n - [H_{\mathrm{LL}}(\beta_n)]^{-1}\nabla_\beta(\beta_n)$$

The concavity of the function guarantees that at every iteration, the parameters produced will increase the log likelihood until the algorithm converges. The Hessian matrix is an estimate of the variance-covariance matrix of components of the parameter vector β.

Thus, if we can show that the Hessian matrix associated with our function is negative definite, then we can apply Newton-Raphson estimation. The Jacobian of the gradient of our log-likelihood function is the Hessian Matrix, which we calculate in Equation 4.2.

$$
\begin{aligned}
H_{\mathrm{LL}}(\beta, N) &= J_\beta \left(\nabla_\beta \mathrm{LL}(\beta)\right) \\
&= \frac{\partial^2}{\partial \beta^2} \mathrm{LL}(\beta) \\
&= \begin{pmatrix} \frac{\partial^2}{\partial \beta_1 \partial \beta_1} \mathrm{LL}(\beta) & \cdots & \frac{\partial^2}{\partial \beta_1 \partial \beta_s} \mathrm{LL}(\beta) \\ \vdots & & \vdots \\ \frac{\partial^2}{\partial \beta_s \partial \beta_1} \mathrm{LL}(\beta) & \cdots & \frac{\partial^2}{\partial \beta_s \partial \beta_s} \mathrm{LL}(\beta) \end{pmatrix} \\
&= \begin{pmatrix} \frac{\partial}{\partial \beta_1} \left(\frac{\partial}{\partial \beta} \mathrm{LL}(\beta)\right)^T \\ \vdots \\ \frac{\partial}{\partial \beta_s} \left(\frac{\partial}{\partial \beta} \mathrm{LL}(\beta)\right)^T \end{pmatrix} \\
&= \frac{\partial}{\partial \beta} \left(\frac{\partial}{\partial \beta} \mathrm{LL}(\beta)\right)^T \\
&= \frac{\partial}{\partial \beta} \left(\nabla_\beta \mathrm{LL}(\beta)\right)^T \\
&= \sum_{n \in N} \sum_{j \in C_n} \frac{1}{N} \frac{\partial}{\partial \beta} \left((z_{nj} - \hat{P}_{nj}(\beta)) X_{nj}\right)^T \\
&= \sum_{n \in N} \sum_{j \in C_n} -\frac{1}{N} \frac{\partial}{\partial \beta} \left(\hat{P}_{nj}(\beta)\right) X_{nj}^T \\
&= \sum_{n \in N} \sum_{j \in C_n} -\frac{1}{N} \left(X_{nj} - \sum_{k \in C_n} X_{nk} \hat{P}_{nk}(\beta)\right) \hat{P}_{nj}(\beta) X_{nj}^T \\
&= \sum_{n \in N} \sum_{j \in C_n} \sum_{k \in C_n} \frac{1}{N} \hat{P}_{nj}(\beta) \hat{P}_{nk}(\beta) X_{nk} X_{nj}^T \\
&\quad - \sum_{n \in N} \sum_{j \in C_n} \frac{1}{N} \hat{P}_{nj}(\beta) X_{nj} X_{nj}^T
\end{aligned} \tag{4.2}
$$

We would like to prove that this matrix is negative definite. In order to do this, we need to show that with any vector a, $a^T H_{\mathrm{LL}}(\beta) a < 0$ for all values of β. We fix n, β, and an arbitrary vector a of length K where K is the number of components of vector β, and

thus the number of columns of X_{nj} for any $j \in C_n$. Let $x_j = a^T X_{nj}$ and $p_j = \hat{P}_{nj}(\beta)$. This means that

$$\sum_{j \in C_n} \sum_{k \in C_n} \hat{P}_{nj}(\beta)\hat{P}_{nk}(\beta) a^T X_{nk} X_{nj}^T a - \sum_{j \in C_n} \hat{P}_{nj}(\beta) a^T X_{nj} X_{nj}^T a$$

$$\begin{aligned}
&= \sum_{j \in C_n} \sum_{k \in C_n} p_j p_k x_j x_k^T - \sum_{j \in C_n} p_j x_j x_j^T \\
&= \sum_{j \in C_n} \sum_{k \in C_n} p_j p_k x_j x_k^T - \sum_{j \in C_n} \sum_{k \in C_n} p_j p_k x_j x_j^T \\
&= \sum_{j \in C_n} \sum_{k \in C_n} p_j p_k (x_j x_k^T - x_j x_j^T) \\
&= \sum_{j<k} p_j p_k (x_j x_k^T - x_j x_j^T + x_k x_j^T - x_k x_k^T) \\
&= \sum_{j<k} -p_j p_k (x_j - x_k)(x_j - x_k)^T
\end{aligned}$$

which is always negative. This means that LL is universally concave, and therefore attains its global maximum at a unique stationary point.

Appendix 3: An introduction to sample weights

Suppose that we have a set U of $|U|$ individuals for whom we wish to know a given quality. If we denote the household by n, let X_n denote the quality of X that we are seeking. This could be response, age, revenue, or any other numeric quantity associated with n. We may wish to know the average value of X_n for the population U. This will be expressed

$$\overline{X_n} = \sum_{n \in U} \frac{X_n}{|U|}$$

However, if we do not have access to the value X_n for all individuals in U, we can estimate $\overline{X_n}$ by taking the weighted average of a sample S of individuals drawn from U. That is, our estimator of $\overline{X_n}$ will be

$$\hat{\overline{X_n}} = \sum_{n \in S} X_n w_n$$

where w_n is the weight of individual X_n chosen in such a way that $\sum_{n \in S} w_n = 1$ and the estimator is unbiased, that is:

$$E(\hat{\overline{X_n}}) = \overline{X_n}$$

We find that

$$\begin{aligned} E(\hat{\overline{X_n}}|T) &= \sum_{n \in T} X_n w_n \\ &= \sum_{n \in U} \delta_{n \in T} X_n w_n \end{aligned}$$

where

$$\delta_A = \begin{cases} 1, \text{ if } A \\ 0, \text{ if not } A \end{cases}$$

If $P(T)$ represents the probability that our sampling strategy will yield the hypothetical sample T, then

$$\begin{aligned} E(\hat{\overline{X_n}}) &= \sum_{T \subset U} \left(\sum_{n \in U} \delta_{n \in T} X_n w_n P(T) \right) \\ &= \sum_{n \in U} X_n w_n \sum_{T \subset U} P(T, n \in T) \\ &= \sum_{n \in U} X_n w_n P(n \in T) \end{aligned}$$

which means that if

$$w_n = \frac{1}{|U| P(n \in T)}$$

then

$$E(\hat{\overline{X_n}}) \;=\; \overline{X_n}$$

and the estimator is unbiased. In simple random sampling, we define the size of the sample we wish to draw from the population, say N_S, and then we set all size-N_S samples of U in the power set of U as being equally likely to be drawn. This means that

$$P(T) \;=\; \begin{cases} k & ,|T| = N_S, T \subset U \\ 0 & , \text{otherwise} \end{cases}$$

So

$$\begin{aligned} 1 &= \sum_{T \subset U} P(T) \\ &= \sum_{|T|=N_S, T \in U} k \\ &= \binom{|U|}{N_S} k \end{aligned}$$

And $P(T) = \binom{|U|}{N_S}^{-1}$. Thus, in simple random sampling,

$$\begin{aligned} P(n \in T) &= \sum_{T \subset U} P(T, n \in T) \\ &= \sum_{|T|=N_S} \binom{|U|}{N_S}^{-1} \delta_{n \in T} \\ &= \binom{|U|-1}{N_S-1} \binom{|U|}{N_S}^{-1} \\ &= \frac{N_S}{|U|} \end{aligned}$$

Finally, we can conclude that

$$\begin{aligned} w_n &= \frac{1}{|U|P(n \in T)} \\ &= \frac{1}{N_S} \end{aligned}$$

In stratified random sampling, which is the sampling strategy that we apply in our case, we divide the population U into a set of strata G, then do simple random sampling within each stratum. Thus, if g is a stratum in G, and U_g is the set of households in U classed in stratum g, N_{U_g} is the size of U_g and N_{S_g} is the number of households that we wish to draw from U_g, then

$$P(n \in T | n \in U_g) = \frac{1}{N_{S_g}}$$

Appendix 4: Entering sample weights in maximum likelihood estimation

In maximum likelihood estimation, we are in effect finding values of β such that the following is true:

$$\nabla_\beta \mathrm{LL}(\beta, S) = \sum_{n \in S} \sum_{j \in C_n} \frac{1}{|S|}(z_{nj} - \hat{P}_{nj}(\beta))X_{nj} \quad = \quad 0 \tag{4.3}$$

which is the estimated covariance between the residuals of the predicted probabilities and the explanatory variables, that we set to zero. We note that if we divide S into sets g of households n and store choices j where both X_{nj} and P_{nj} are invariant (as is the case for households within the same iris in our survey), we can define X_g as the value X_{nj} attains for households and store choices in g, and P_g is the value that P_{nj} attains for households and store choices in g. Then, if $|S_g|$ is the number of household-store choice pairs in g that appear in the sample, and

$$\bar{z}_g \quad = \quad \sum_{(n,j) \in g} \frac{z_{nj}}{|S_g|}$$

then

$$\nabla_\beta \mathrm{LL}(\beta, S) = \sum_{g \in G} \sum_{(n,j) \in g} \frac{1}{|S_g|} N_g (\bar{z}_g - \hat{P}_g(\beta)) X_g$$

The log likelihood estimate is a measure of the distance between the empirical probabilities $\bar{z}_g$ of selecting a given alternative, given a given set of alternatives, and the predicted probabilities associated with these values, weighted by the number of observations in our data set corresponding to these observed values. If we introduce our sampling weights, we can replace the empirical probability $\bar{z}_g$ of households in group g selecting the given alternative with the an estimate of the proportion of households within this group in the general population selecting the given alternative. We do this by replacing $\bar{z}_g$ with

$$\bar{z}_{g,w} = \frac{\sum_{(n,j) \in g} w_n z_{nj}}{N_g}$$

where N_g is the number of household-store choice pairs in g that appear in the general population. Then, if we define

$$w_g = \sum_{(n,j) \in g} \frac{w_n}{N}$$

we have the formula:

$$\nabla_\beta \mathrm{LL}(\beta, S) = \sum_{g \in G} \sum_{(n,j) \in g} w_g (\bar{z}_{g,w} - \hat{P}_g(\beta)) X_g$$

Appendix 5: Development of the gravitational model

In this appendix, we show the various ways that we looked at estimating the parameters of our gravitational model. In all, we have three techniques that we shall describe, the last technique being only a very slight variation on the technique that we chose to use in our model.

We begin by showing why we choose to remove the retail space term from our gravitational model. Once we take into account the store type of the store in the model, the retail space of the store has little effect on determining the likelihood of choosing a store, since an important attribute of store type is the general size of the retail space. We confirm this with a quick initial investigation. We look at all pairs of households and large-surface stores and group them by the distances between store and domicile, and by retail space. Household-store-choice pairs are grouped in intervals of 1000 meters. Thus, the first class contains store choices that are less than 1 km away from the store. The next, between 1000 and 2000 meters away, and so on. The intervals defined for retail space classes have to vary in order to ensure that there are no intervals defined that represent no household-store choice pair. Stores that have less than 2500 square meters of retail space are grouped in intervals of 100 square meters. Stores between 2500 and

12,000 square meters are grouped in intervals of 500 square meters. All other stores fall under categories of greater than 12,000 square meters.

We use these categories of distance and surface areas in order to divide the set of all household-store-choice pairs into classes. These classes will be defined in three ways, either by the categories of distance only, by the categories of store retail space, or by the categories of both distance and retail space. Within each of these classes, we shall calculate the percent of each household-store-choice pairs representing an observed store choice. We shall then use linear regression to see if we can see a relation between the log of the average distance represented in each group, and the log of the average retail space represented in each group.

Thus, for example, in order to find values of $\alpha(\text{SM})$ and $\beta(\text{SM})$, we use three different models,

$$\begin{aligned} \log Y_{g_d}(\text{SM}) &= K + \beta(\text{SM}) \log d_{g_d}(\text{SM}) \\ \log Y_{g_s}(\text{SM}) &= K + \alpha(\text{SM}) \log s_{g_s}(\text{SM}) \\ \log Y_{g_{sd}}(\text{SM}) &= K + \alpha(\text{SM}) \log s_{g_{sd}}(\text{SM}) + \beta(\text{SM}) \log d_{g_{sd}}(\text{SM}) \end{aligned} \tag{4.4}$$

where g_d defines the set of household-store choice pairs corresponding to distances within the set d, g_s defines the set of household-store choice pairs corresponding to retail spaces within the interval s, and g_{sd} corresponds to household-store choice pairs with distances within the interval d and retail spaces within the interval s. Y_g is the proportion of household-store choice pairs in the group g that correspond to an observed store choice, and d_g corresponds to the average distance represented by household-store choice pairs

in g, and s_g represents the average retail space for stores in household-store choice pairs in g. K is an intercept term. We use a simple linear regression to calculate each of these models, in each case eliminating any groups g for which the value of Y is zero.

We present the R^2 terms of these models in the Table 4.21:

Store Type	Dist Only	Surf Only	Dist,Surf
SM	0.863	0.266	0.654
HM	0.827	0.019	0.558
HD	0.808	0.242	0.343
XM	0.745	0.203	0.239

Table 4.21: R^2 terms for regression models of gravitational parameters.

We can see that indeed, the fit of a model not including retail space is much better than one that does include it. The coefficients of the various models are in Table 4.22:

Model	Dist Only	Surf Only	Dist,Surf	
Coefficient	Distance	Retail Space	Distance	Retail Space
SM	−1.768	−6.225	−1.279	1.174
HM	−1.818	−0.991	−1.409	0.903
HD	−1.663	−6.322	−0.540	0.692
XM	−1.503	−3.564	−0.703	2.000

Table 4.22: Parameter estimates of regression model of gravitational parameters.

We find that the retail space is nonsignificant in the model of the percent of stores chosen in groups defined only be retail space, but it is significant and more representative of what we would expect when entered in a model of percent of stores chosen in groups defined by retail space and distance. However, as we saw by the R-squared terms, the model is weakened when we take retail space into account. The greater variation in the Y terms introduced by the increased number of classes of household-store choice pairs, and

the smaller number of observations in each class could in effect offset any improvements in our model brought about by having more explanatory variables.

This can be shown in the scatterplots of Figures 4.6 and4.7.

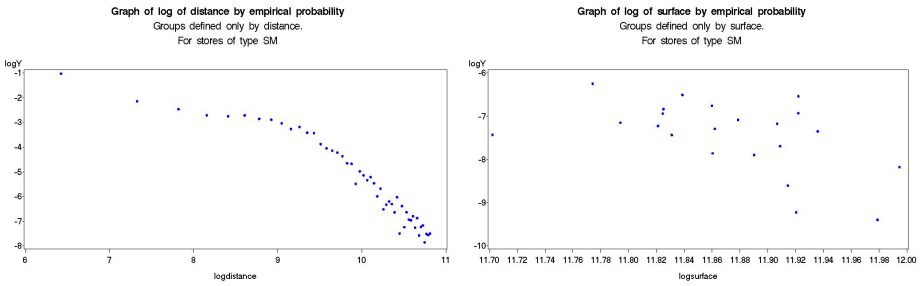

Figure 4.6: Scatterplots of log of distance and log of retail space (surface) and proportion choosing store for model based on distance, and retail space, respectively.

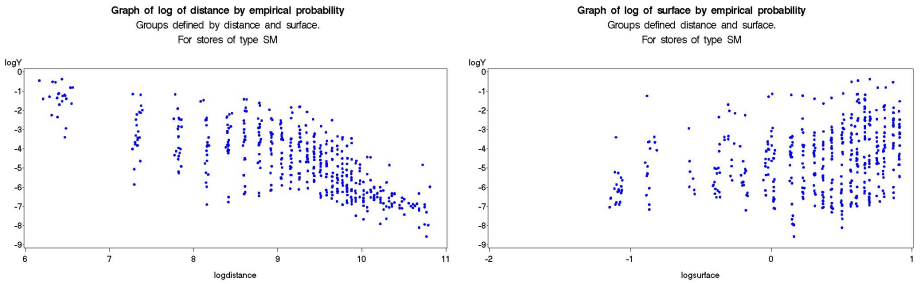

Figure 4.7: Scatterplots of log of distance and log of retail space (surface) and proportion choosing store for model based on both distance and retail space.

We have therefore rewritten the model utility in the following manner:

$$A_{nj} \quad = \quad d_{nj}^{\beta(t(j))}$$

This leaves us with four parameters that we need to find: β(sm), β(hm), β(hd), and

β(xm) (corresponding, of course, to supermarkets, hypermarkets, hard discount stores and large hypermarkets). We estimate each of these parameters independently, adapting them to the population of households observed selecting the outside option and observed selecting a store that is of the corresponding store type. We have derived three techniques in order to find the β parameters of our model. For each of these, we will suppose that we are trying to find the parameter β(sm), since the estimation of the other parameters will be identical to the first.

Technique 1

The first technique that we use is to treat our gravitational model as we would a conditional logit model. That is, the probability that individual n chooses j conditional on the individual choosing a store that is of store type sm will be

$$\begin{aligned} P_{nj|\text{sm}} &= \frac{d_{nj}^{\beta(\text{sm})}}{\sum_{k \in C_{n|\text{sm}}} d_{nk}^{\beta(\text{sm})}} \\ &= \frac{\beta(\text{sm}) \ln d_{nj}}{\sum_{k \in C_{n|\text{sm}}} \beta(\text{sm}) \ln d_{nk}} \end{aligned}$$

where $C_{n|\text{sm}}$ is the set of alternatives in the choice set of n corresponding to supermarkets. We recall that to do maximum likelihood estimation of β(sm), we find the value of β that maximizes the expression

$$\text{LL}(\beta) = \sum_{n=1}^{N} \sum_{j \in J_{n,\text{sm}}} d_{nj} \ln P_{nj|\text{sm}}(\beta)$$

where d_{nj} is one when household n chooses j and zero, otherwise. We recall here that due to the nature of our data collection, household co-ordinates were assigned the centroid of the "IRIS" in which they resided, implying that all households in the same "IRIS" are

assigned the same set of stores in their choice set, and were assigned the same distances between their homes and their stores. This will allow us to simplify our formula. If we create a partition G of the sample S such that for any group g in G, if $n \in g$, $C_{n,\text{sm}} = C_g$, and $d_{nk} = d_{gk} \forall k \in C_g$, then

$$\begin{aligned} \text{LL}(\beta) &= \sum_{g \in G} \sum_{n \in g} \sum_{j \in C_g} d_{nj} \ln P_{nj|sm}(\beta) \\ &= \sum_{g \in G} \sum_{j \in C_g} w_{gj} \ln P_{gj|sm}(\beta) \\ &= \sum_{g \in G} \sum_{j \in C_g} w_{gj} \ln \left(\frac{d_{nj}^{\beta}}{\sum_{k \in C_{n|\text{sm}}} d_{nk}^{\beta}} \right) \end{aligned}$$

for $w_g = \sum_{n \in g} d_{nj}$. This is a simpler equation to maximize than the original maximum likelihood estimate formula, however, it is not simple enough for our purposes. In our data, the number of points to sum is reduced by an order of ten, but there still remain over 2 million observations on which to calculate, and so we seek a simpler technique.

Technique 2

This technique finds the values of the parameters that minimize the misspecification error of our model. Recall that our model took the form

$$P_{nj} = \frac{e^{\sum_{t \in T} \beta_t X_{njt}}}{\sum_{k \in C_n} e^{\sum_{t \in T} \beta_t X_{nkt}}}$$

when we supposed that the set of variables $\{X_{nj1}, \ldots, X_{njT}\}$ represented the explanatory variables of a fully specified model. In order to save time, we can remove all the explanatory variables from the model but the variable representing distance, letting

$X_{nj1} = \ln d_{nj}$, so that

$$P_{nj} = \frac{e^{\beta \ln(d_{nj})}}{\sum_{k \in C_n} e^{\beta \ln(d_{nk})}}$$

However, this will represent a grossly underspecified model, resulting in the estimation of the coefficient of the distance term that is influenced by effects other than distance and thus biased. Instead of making the false assumption that the probabilities of selection are entirely determined by distance, we represent all the omitted explanatory variables representing effects associated with household n and individual j that are independent of distance with the random variable δ_{nj}. It would not be unreasonable to suppose this variable to be normally distributed, since it is a sum of independent effects. We can also assume that the characteristics of each store are independent of each other (meaning we ignore the effect of spatial autocorrelation, among other things), resulting in these random variables being independent. We also assume, for the sake of mathematical convenience, that this variable is centred and homoskedastic, with variance σ^2. Our function then becomes:

$$P_{nj} = \frac{e^{\beta \ln(d_{nj}) + \delta_{nj}}}{\sum_{k \in C_n} e^{\beta \ln(d_{nk}) + \delta_{nk}}}$$

This equation can be linearized, in order to allow for least squares estimation. We begin by taking the log of both sides of the equation to get:

$$\ln P_{nj} = \beta \ln d_{nj} + \delta_{nj} - \ln \left(\sum_{k \in C_n} e^{\beta \ln(d_{nk}) + \delta_{nk}} \right)$$

We then take the average of both sides of the equation for alternatives in the choice set

C_n of n. If N_{C_n} is the cardinality of the set C_n then:

$$\frac{1}{N_{C_n}} \sum_{j \in C_n} \ln P_{nj} \quad = \quad \frac{1}{N_{C_n}} \sum_{j \in C_n} \left[\beta \ln d_{nj} + \delta_{nj} - \left(\sum_{k \in C_n} e^{\beta \ln(d_{nk}) + \delta_{nk}} \right) \right]$$

Isolating the last term, we obtain:

$$\frac{1}{N_{C_n}} \sum_{j \in C_n} \left(\sum_{k \in C_n} e^{\beta \ln(d_{nk}) + \delta_{nk}} \right) \quad = \quad \frac{1}{N_{C_n}} \sum_{j \in C_n} [\beta \ln d_{nj} + \delta_{nj} - \ln P_{nj}]$$

which we can replace in the equation above to obtain:

$$\ln P_{nj} - \sum_{k \in C_n} \frac{\ln P_{nk}}{N_{C_n}} \quad = \quad \beta \ln d_{nj} - \beta \sum_{k \in C_n} \frac{\ln d_{nk}}{N_{C_n}} + \delta_{nj} - \sum_{k \in C_n} \frac{\delta_{nk}}{N_{C_n}}$$

This can be expressed in matrix notation, where we let

$$\mathbf{P}_n = \begin{pmatrix} \vdots \\ \ln P_{nk} \\ \vdots \end{pmatrix}, \mathbf{d}_n = \begin{pmatrix} \vdots \\ \ln d_{nk} \\ \vdots \end{pmatrix}, \text{ and } \delta_n = \begin{pmatrix} \vdots \\ \delta_{nk} \\ \vdots \end{pmatrix}$$

For any individual, then, the equation above becomes:

$$\left(\mathbf{I}_n - \frac{1}{N_{C_n}} \mathbf{U}_n \right) \mathbf{P}_n \quad = \quad \beta \left(\mathbf{I}_n - \frac{1}{N_{C_n}} \mathbf{U}_n \right) \mathbf{d}_n + \left(\mathbf{I}_n - \frac{1}{N_{C_n}} \mathbf{U}_n \right) \delta_n$$

where $\mathbf{I}_n$ is the identity matrix of dimension n, and $\mathbf{U}_n$ is the unit matrix of dimension n containing ones in all rows and columns. This, we can simplify by introducing the notation $\Delta_n = \left(\mathbf{I}_n - \frac{1}{N_{C_n}} \mathbf{U}_n \right)$ so that

$$\Delta_n \mathbf{P}_n \quad = \quad \beta \Delta_n \mathbf{d}_n + \Delta_n \delta_n \tag{4.5}$$

We note that $\Delta_n^T = \Delta_n$ and that $\Delta_n\Delta_n = \Delta_n$ meaning that this matrix is not invertible.

By the supposition of our model,

$$\delta_n \sim \mathrm{N}\left(\mathbf{0}_n, \sigma^2\mathbf{I}_n\right)$$

and therefore, if $\varepsilon_n = \Delta_n\delta_n$, then

$$\varepsilon_n \sim \mathrm{N}\left(\mathbf{0}, \sigma^2\Delta_n\right)$$

For us to be able to use GLS estimation, the covariance matrix of our error terms must be invertible. However, that is not the case. The last row of the Δ_n matrix is equal to the negative of the sum of all the other rows in the matrix. This means that if we define the $N_{C_n} - 1$ by N_{C_n} matrix

$$\mathbf{R}_n = \begin{pmatrix} 1 & 0 & \dots & 0 & -1 \\ 0 & 1 & \dots & 0 & -1 \\ & & \dots & & \\ 0 & 0 & \dots & 1 & -1 \end{pmatrix} \tag{4.6}$$

then estimating β using the equation

$$\mathbf{R}_n\Delta_n\mathbf{P}_n = \beta\mathbf{R}_n\Delta_n\mathbf{d}_n + \mathbf{R}_n\Delta_n\delta_n$$

is equivalent to using Equation 4.5. Since $\mathbf{R}_n\Delta_n = \mathbf{R}_n$, this becomes

$$\mathbf{R}_n\mathbf{P}_n = \beta\mathbf{R}_n\mathbf{d}_n + \mathbf{R}_n\delta_n$$

Now

$$\gamma_n = \mathbf{R}_n\delta_n \sim \mathrm{N}\left(\mathbf{0}, \sigma^2\mathbf{R}_n\mathbf{R}_n^T\right)$$

and the covariance matrix $\mathbf{R}_n\mathbf{R}_n^T$ is invertible so the following formula will give us the GLS estimate of β:

$$\mathrm{GLS}(\beta) = \sum_{n\in N}\Big((\mathbf{R}_n\mathbf{d}_n)^T\left(\mathbf{R}_n\mathbf{R}_n^T\right)^{-1}(\mathbf{R}_n\mathbf{d}_n)\Big)^{-1}(\mathbf{R}_n\mathbf{d}_n)^T\left(\mathbf{R}_n\mathbf{R}_n^T\right)^{-1}(\mathbf{R}_n\mathbf{P}_n)$$

This can be further simplified by looking at the partition G of the sample N where C_n and d_{nj} are invariant for all n in g. Thus,

$$\begin{aligned}
\mathrm{GLS}(\beta) &= \sum_{n\in N}\Big((\mathbf{R}_n\mathbf{d}_n)^T\left(\mathbf{R}_n\mathbf{R}_n^T\right)^{-1}(\mathbf{R}_n\mathbf{d}_n)\Big)^{-1}(\mathbf{R}_n\mathbf{d}_n)^T\left(\mathbf{R}_n\mathbf{R}_n^T\right)^{-1}(\mathbf{R}_n\mathbf{P}_n)\\
&= \sum_{g\in G}\sum_{n\in g}\Big((\mathbf{R}_n\mathbf{d}_n)^T\left(\mathbf{R}_n\mathbf{R}_n^T\right)^{-1}(\mathbf{R}_n\mathbf{d}_n)\Big)^{-1}(\mathbf{R}_n\mathbf{d}_n)^T\left(\mathbf{R}_n\mathbf{R}_n^T\right)^{-1}(\mathbf{R}_n\mathbf{P}_n)\\
&= \sum_{g\in G}\sum_{n\in g}\Big((\mathbf{R}_g\mathbf{d}_g)^T\left(\mathbf{R}_g\mathbf{R}_g^T\right)^{-1}(\mathbf{R}_g\mathbf{d}_g)\Big)^{-1}(\mathbf{R}_g\mathbf{d}_g)^T\left(\mathbf{R}_g\mathbf{R}_g^T\right)^{-1}(\mathbf{R}_g\mathbf{P}_g)\\
&= \sum_{g\in G}|g|\Big((\mathbf{R}_g\mathbf{d}_g)^T\left(\mathbf{R}_g\mathbf{R}_g^T\right)^{-1}(\mathbf{R}_g\mathbf{d}_g)\Big)^{-1}(\mathbf{R}_g\mathbf{d}_g)^T\left(\mathbf{R}_g\mathbf{R}_g^T\right)^{-1}(\mathbf{R}_g\mathbf{P}_g)
\end{aligned}$$

where obviously N_g is the number of individuals in g. Since we do not have the value of P_{gj}, equal to P_{nj} for all n in g, we replace it with the estimator

$$O_{gj} = \sum_{n\in g}\frac{O_{nj}}{N_g} \tag{4.7}$$

where O_{nj} is one when n chooses j and zero, otherwise. O_{gj} is an unbiased estimator of

P_{nj} for any individual n in N_g, since

$$\begin{aligned} E\left(O_{gj}\right) &= E\left(\sum_{n\in g}\frac{O_{nj}}{N_g}\right) \\ &= \sum_{n\in g}\frac{P_{nj}}{N_g} \\ &= \sum_{n\in g}\frac{P_{gj}}{N_g} \\ &= P_{gj} \end{aligned} \tag{4.8}$$

We can then calculate the vector

$$\mathbf{O}_g = \begin{pmatrix} \cdots \\ \ln O_{gk} \\ \cdots \end{pmatrix}$$

so that

$$\hat{\mathrm{GLS}}(\beta) = \sum_{g\in G} N_g \left(\left(\mathbf{R}_g\mathbf{d}_g\right)^T\left(\mathbf{R}_g\mathbf{R}_g^T\right)^{-1}\left(\mathbf{R}_g\mathbf{d}_g\right)\right)^{-1}\left(\mathbf{R}_g\mathbf{d}_g\right)^T\left(\mathbf{R}_g\mathbf{R}_g^T\right)^{-1}\left(\mathbf{R}_g\mathbf{O}_g\right)$$

We note that the measured distance between household and store, d_{nj} is always nonzero, and P_{nj} is necessarily so, since it is a function of exponential functions. Therefore, we can be assured that the vectors $\mathbf{d}_g$ and $\mathbf{P}_g$ are always well-defined. Unfortunately, this is not the case for $\mathbf{O}_g$, since we have found that the measured values O_{gj} are frequently zero, rendering many vector compenents $\ln O_{gj}$ undefined. In fact, due to insufficiently large sample sizes, and because most stores in each choice set are far from the household and therefore have extremely low probabilities of being selected, the vast

majority of these measured empirical probabilities of selection are zero, so this is not a problem we can ignore.

We emphasize here that since our original model equation allows for zero values of P_{nj}, we could use maximum likelihood estimation of our parameters in the equation

$$O_{gj} = \frac{e^{\beta \ln(d_{gj})+\delta_{gj}}}{\sum_{k\in C_g} e^{\beta \ln(d_{gk})+\delta_{gk}}}$$

It is in fact, simply the linearization of our expression through the use of logarithms that creates this problem.

In order to render our expression well-defined, therefore, we replace the vector $\mathbf{O}_g$ with $\tilde{\mathbf{O}}_g$ where the latter vector is simply the vector $\mathbf{O}_g$ with all undefined components corresponding to values of O_{gj} equal to zero removed. If $\mathbf{E}_g$ is the identity matrix of dimension equal to N_{C_g} with all rows corresponding to undefined components of $\mathbf{O}_g$ eliminated, then

$$\tilde{\mathbf{O}}_g = \mathbf{E}_g \mathbf{O}_g$$

By multiplying $\mathbf{E}_n$ (equal to $\mathbf{E}_g$ for individuals n in group g) with both sides of equation 4.5, we obtain

$$\mathbf{E}_n \Delta_n \mathbf{P}_n = \beta \mathbf{E}_n \Delta_n \mathbf{d}_n + \mathbf{E}_n \Delta_n \delta_n \tag{4.9}$$

We note that as long as $\mathbf{E}_g \Delta_g \mathbf{E}_g^T$ has rank less than N_g we can assume that it is invertible, and we don't need to multiply by the matrix R_g. From this modified equation, the

estimator $\hat{\text{GLS}}$ becomes:

$$\hat{\text{GLS}}(\beta) = \sum_{g \in G} N_g \quad \left((\mathbf{E}_g \Delta_g \mathbf{d}_g)^T \left(\mathbf{E}_g \Delta_g \mathbf{E}_g^T \right)^{-1} (\mathbf{E}_g \Delta_g \mathbf{d}_g) \right)^{-1} \times (\mathbf{E}_g \Delta_g \mathbf{d}_g)^T \left(\mathbf{E}_g \Delta_g \mathbf{E}_g^T \right)^{-1} (\mathbf{E}_g \Delta_g \mathbf{O}_g)$$

which is a well-defined expression.

This solution is not quite satisfactory, since we expect the results to be biased by the fact that by eliminating observations where O_{gj} was zero, we were in effect eliminating a disproportionate number of observations where P_{gj} was small. However, the determination of this bias is extremely difficult, since the construction of the matrix E_g depends on random effects highly correlated with O_{gj}.

We use this estimation technique to obtain the parameter estimates in Table 4.23:

Supermarkets		Hypermarkets		Hard Discount		L. Hypermarkets	
β	R^2	β	R^2	β	R^2	β	R^2
-0.458	0.346	-0.348	0.220	-0.248	0.174	-0.842	0.441

Table 4.23: Gravitational parameter estimates using "Technique 2" for the estimation.

We see here that our regression produces very poor R-squared terms, indicating that our regression does not fit well with the data. As well, we expect that stores are less attractive as they become further away, and yet we do not always produce negative parameter estimates, especially when we calculate our parameter estimates on models that include fewer outside options. We believe that the poor model performance likely comes from the amount of uncertainty entered into our model due to the number of empirical probabilities we must calculate on the basis of very few observed store choices. This can be illustrated in the scatterplots in Figures 4.8 and 4.9. On the y-axis, we show

the values

$$\ln O_{gj} - \frac{1}{N_g} \sum_{k \in C_g} \ln O_{gk}$$

where $\mathbf{O}_{gj}$ is nonzero, that we calculate for each group g and store choice j when we considered all large-surface stores as belonging to the outside set. On the x-axis, we show the values

$$\ln d_{gj} - \frac{1}{N_g} \sum_{k \in C_g} \ln d_{gk}$$

Our scatterplots seem to show that any relationship between these two values is dominated by random effects, and as we have seen above, our model does not take into account the complexity of the relationship between these two values, and effects of bias are unquantified.

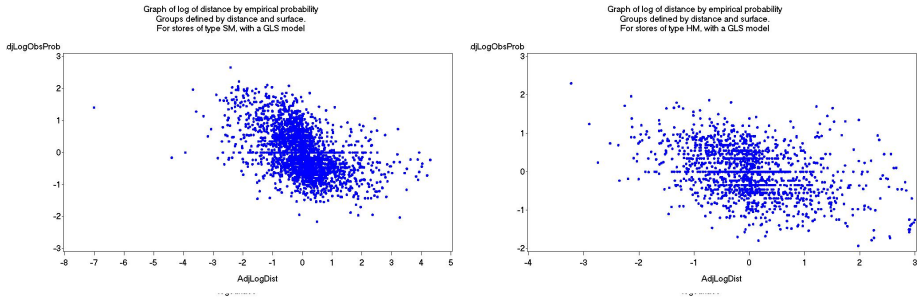

Figure 4.8: Scatterplots of log of distance and log of the empirical probabilities of selecting stores for supermarkets, and small hypermarkets.

Technique 3

The last technique that we develop makes use of the assumption that the total sum of all the weights of all household-store choice pairs is roughly constant. We define the

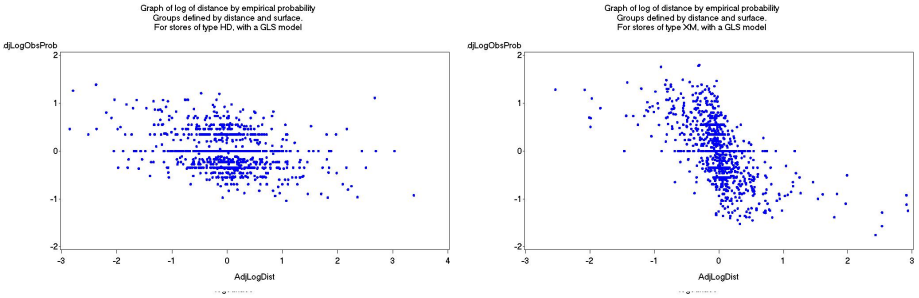

Figure 4.9: Scatterplots of log of distance and log of the empirical probabilities of selecting stores for hard discount stores, and large hypermarkets.

probability of selecting a given store as follows:

$$\begin{aligned} P_{nj} &= \frac{d_{nj}^{\beta}}{\sum_{k \in C_n} d_{nk}^{\beta}} \\ &= \frac{d_{nj}^{\beta}}{d_{nj}^{\beta} + K} \end{aligned}$$

which becomes

$$\frac{K P_{nj}}{1 - P_{nj}} = d_{nj}^{\beta} \tag{4.10}$$

Taking logs of both sides:

$$\ln\left(\frac{P_{nj}}{1 - P_{nj}}\right) = \alpha + \beta \ln d_{nj}$$

We then partition all pairs of households and store choices (n, j) into a set of subgroups G such that for all pairs (n, j) in g, $d_{nj} \approx d_g$. We define

$$P_g \approx \frac{\sum_{(n,j) \in g} P_{nj}}{N_g}$$

and

$$O_g \approx \frac{\sum_{(n,j)\in g} O_{nj}}{N_g}$$

Provided that we have well-defined the groups g, for all (n, j) in g,

$$\alpha + \beta \ln d_{nj} = \ln\left(\frac{P_{nj}}{1 - P_{nj}}\right) \approx \ln\left(\frac{P_g}{1 - P_g}\right) = \alpha + \beta \ln d_g$$

we shall assume that this holds as a strict equality and that therefore P_{nj} is strictly equal to P_g, provided (n, j) is contained in g. Let

$$O_g = P_g + \delta_g$$

Since $\mathrm{E}(O_g) = P_g$, and $\mathrm{E}(\delta_g) = 0$, and since $\mathrm{var}(O_g) = \frac{P_g(1-P_g)}{N_g}$, this will be the variance of δ_g. We shall assume these error terms to be independent of the error terms in other groups g. We can now derive a formula, based on the empirical probabilities of selection that is a linear expression of the model parameters:

$$\begin{aligned} \ln\left(\frac{O_g}{1 - O_g}\right) &= \ln\left(\frac{P_g + \delta_g}{1 - P_g - \delta_g}\right) \\ &= \ln(P_g + \delta_g) - \ln(1 - P_g - \delta_g) \end{aligned}$$

Which, if we linearize with a partial Taylor expansion, obtains

$$\begin{aligned} \ln\left(\frac{O_g}{1 - O_g}\right) &\approx \ln(P_g) + \frac{\delta_g}{P_g} - \ln(1 - P_g - \delta_g) + \frac{\delta_g}{1 - P_g} \\ &= \ln\left(\frac{P_g}{1 - P_g}\right) + \frac{\delta_g}{P_g(1 - P_g)} \end{aligned}$$

We replace $\ln\left(\frac{P_g}{1-P_g}\right)$ with $\alpha + \beta \ln d_g$ and $\frac{\delta_g}{1-P_g}$ with ε_g to obtain

$$\ln\left(\frac{O_g}{1-O_g}\right) \approx \alpha + \beta \ln d_g + \varepsilon_g \tag{4.11}$$

With a correction for heteroskedacity, this equation fulfills the conditions for least squares estimation, which we can obtain by minimizing the following formula:

$$\sum_{g \in G} w_g \left(\ln\left(\frac{O_g}{1-O_g}\right) - \alpha - \beta \ln d_g\right)^2$$

Where

$$w_g = (N_g(O_g(1-O_g)))^{-\frac{1}{2}}$$

which is the inverse of the variance of ε_g, with the value P_g estimated with O_g. We note that O_g must be nonzero for all values of g. This strategy has the advantage over the use of the formula $P_{nj} = \alpha + \beta_j d_{nj}$ that we use in our thesis in that it explicitly defines a probability that is bounded between 0 and 1. However, our work with this technique did not yield results that were better than the technique that we chose.

In our thesis, we have used ordinary linear regression on empirical probabilities of selection on small subsamples of our population in order to do imputation. However, this was not before considering a slightly different type of regression. Dividing up the household-store-choice pairs by distance, as we did in 3.2, we obtain the scatterplots in Figures 4.10 to 4.13.

The scatterplots show that the relationship between the log of the distance and the log of the observed proportion of households selecting the store within each class of

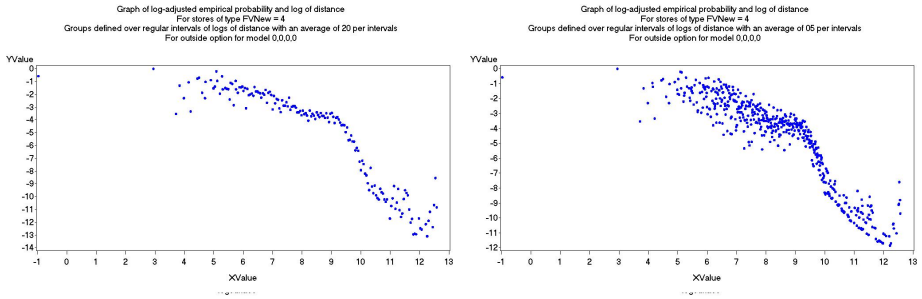

Figure 4.10: Comparison of scatterplots of log of distances of supermarkets and log of proportion corresponding to observation of supermarket choice for K equal to 20 and K equal to 5.

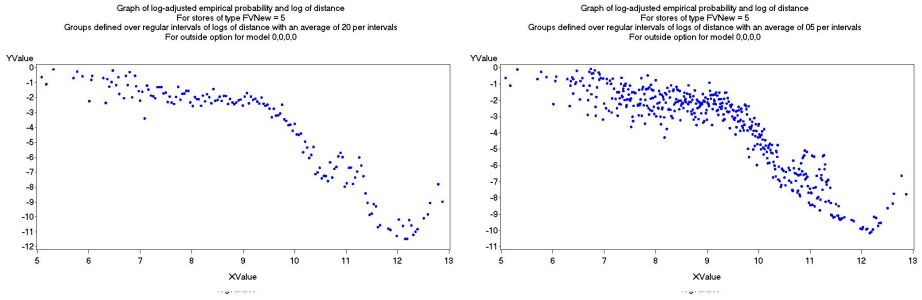

Figure 4.11: Comparison of scatterplots of log of distances of small hypermarkets and log of proportion corresponding to observation of small hypermarket choicefor K equal to 20 and K equal to 5.

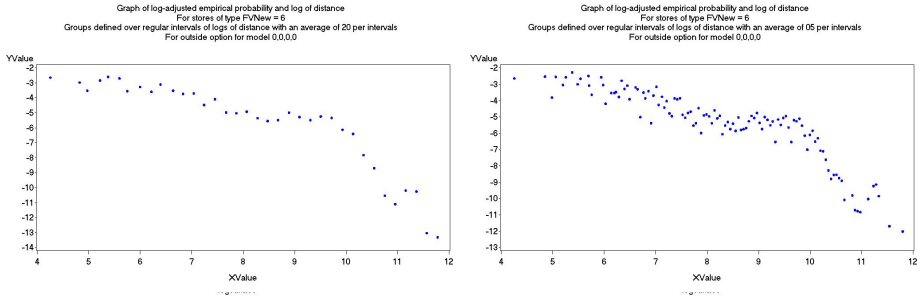

Figure 4.12: Comparison of scatterplots of log of distances of large hypermarkets and log of proportion corresponding to observation of large hypermarket choice for K equal to 20 and K equal to 5.

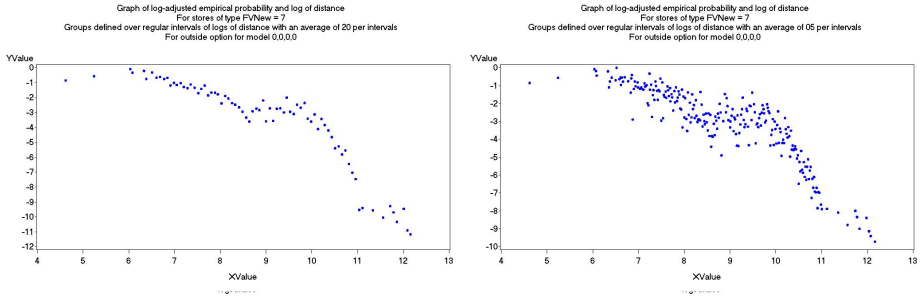

Figure 4.13: Comparison of scatterplots of log of distances of hard discounts and log of proportion corresponding to observation of hard discount choice for K equal to 20 and K equal to 5.

distance is not linear. It would seem that the scatterplots would be best fit with a continuous function that has a nondifferentiably point at roughly 10,000 meters ($\approx e^{9.2}$) for supermarkets and small hypermarkets, and at roughly 15,000 meters for hard discount stores and large hypermarkets. It seems that for hard discounts and large hypermarkets, there may be a much fainter second nondifferentiable point at roughly $e^{8.5}$ or 5000 meters. That is, a household's propensity to choose a hard discount store and a large hypermarket remains unchanged as the distance goes from 5000 meters to 1500 meters, but as the distance increases further, the propensity drops. This scatterplot pattern remains the same for all choices of large-surface store.

Our model is intended to be a very rough approximation of the relationship between distance and the tendency to select a large-surface store, and because we have a much more sophisiticate logit model that provides better predictions of a household's likelihood of choosing a given alternative, seeking corrections to our linear estimation that provide a better fit are unnecessary. However, we chose to verify the improvements a very slight modification of our linear regression would make to our model fit.

Instead of fitting one single straight line to our data, we fit two lines, so that we have one linear regression that applies to stores that correspond to distances lower than the

one found at the cusp of the curve, and another that applied to other distances. There are statistical procedures that can be used to estimate at what point the linear coefficient of regression changes. Krieger, Pollak, and Yakir (2003, [22]), for example, describe how Cumulative Sum (CuSum) procedures can be used in linear regression in order to find the point at which the regression changes. We did not do this. Instead, since the cusps swere rather easy to see on the scatterplots, we found them through visual inspection. What we found was that when we used a gravitational model where there were two slope parameters of regression, the quality of the results were not any better than those that were obtained through a gravitational model whose parameters were estimated with a straightforward linear regression.

Appendix 6: Survey questionnaire

The following pages contain the first few pages extracted from the survey questionnaire from BVA's study of household shopping flux in the Centre Region which concern questions about food products purchased in large-surface stores.

DPT

JOUR DE REALISATION

HEURE DE REALISATION

TELEPHONE

INTER

VRAIINSEE

N°ADRESSE BVA

ETAT

- ❑ PAS DE TONALITE
- ❑ OCCUPE
- ❑ NRP
- ❑ A RACCROCHE
- ❑ REPONDEUR
- ❑ PERSONNE ABSENTE DUREE ENQUETE
- ❑ REFUS
- ❑ FAUX NUMERO
- ❑ RENDEZ-VOUS
- ❑ HORS QUOTAS
- ❑ FAX / MODEM
- ❑ DOUBLON
- ❑ HORS CIBLE
- ❑ NE PEUT COMMUNIQUER
- ❑ DECONNECTION CLIENT
- ❑ INCONNU

- ❑ ABANDONS
- ❑ HORS QUOTAS 2
- ❑ PAS UTILISE
- ❑ PAS UTILISE 2
- ❑ PAS UTILISE 3

CONTACT

NOM

PRENOM

ADR1

ADR2

ADR3

COMMUNE

CODE POSTAL

NOM_IRIS

SEED

Si vrai, spécifier la valeur '??SEED??' à la question 'identifiant'

DPT

REPRISEAFFICHAGE

- ❑ BOURGES
- ❑ CHARTRES
- ❑ TOURS
- ❑ CHATEAUROUX
- ❑ BLOIS
- ❑ ORLEANS

SECT QUOTAS NVX IMPORTS

- ❑ 1801
- ❑ 1802
- ❑ 1803
- ❑ 1804
- ❑ 1805
- ❑ etc

ENQUETEUR VOUS APPELEZ :

Téléphone : TEL

CP : CODE POSTAL

VILLE : VILLE

DEPARTEMENT : DPT

- ❑ Bon numéro

Bonjour, je suis NOMENQUETEUR de la société BVA.

Nous réalisons une enquête concernant les habitudes de consommation. Cette enquête est demandée par les Chambres de Commerce de REPRISEAFFICHAGE. Cette étude vise à savoir où vous achetez différents types de produits. Nous cherchons à mieux connaître le fonctionnement du commerce dans le département pour essayer d'adapter au mieux ses évolutions aux besoins et aux pratiques des consommateurs.

Votre foyer a été sélectionné pour répondre en famille à un questionnaire par lequel nous souhaitons connaître, c'est à dire les magasins que vous fréquentez pour vos achats de produits alimentaires et non alimentaires ainsi que quelques informations sur votre foyer.

L'enquête dure environ 30 minutes. Je souhaite interroger la personne qui fait habituellement les achats pour le foyer. Accepteriez-vous de participer à cette enquête ?

Accepteriez-vous de répondre à quelques questions ?

ENQUETEUR : SI LA PERSONNE N'EST PAS DISPONIBLE, PRENDRE RENDEZ-VOUS

- ❑ Oui, je réponds de suite
- ❑ Non cela ne m'intéresse pas.
- ❑ Non, je n'ai pas le temps
- ❑ Non, je ne veux pas donner d'informations sur mon foyer
- ❑ A quoi cela sert ? qui a commandé l'étude
- ❑ Qu'est-ce que cela va me rapporter

CHAP FILTRE

Vous savez, cette enquête concerne l'avenir de votre région et à ce titre votre participation, votre opinion en tant que consommateurs nous sont très précieuses.

- ❑ VALIDATION
- ❑ REFUS

Si = 1, aller à 'INTRO'

Le questionnaire dure environ 30 minutes, je peux vous rappeler à un autre moment

- ❑ VALIDATION
- ❑ REFUS

Si = 1, aller à 'INTRO'

Vos réponses resteront confidentielles et serviront à l'établissement de statistiques anonymes, des moyennes, des tableaux, des graphiques..

Aucun lien entre votre identité et vos réponses ne sera établi.

- ❑ VALIDATION
- ❑ REFUS

Si = 1, aller à 'INTRO'

L'étude est demandée par la Chambre de Commerce de REPRISEAFFICHAGE

L'étude sert à mieux comprendre comment fonctionne le commerce pour essayer de l'améliorer,

notamment pour savoir où il faut implanter des grandes surfaces et pour essayer de maintenir des petits commerces là où les gens en ont besoin.

- ❑ VALIDATION
- ❑ REFUS

Si = 1, aller à 'INTRO'

L'étude est demandée par la Chambre de Commerce de REPRISEAFFICHAGE, qui est un organisme public.

La Chambre de commerce cherche à mieux connaître les comportements des consommateurs, pour mieux adapter le commerce à vos besoins.

- ❑ VALIDATION
- ❑ REFUS

Si = 1, aller à 'INTRO'

Car1: Je vais tout d'abord vous poser quelques questions de caractéristiques.

Etes-vous célibataire, ou en couple ?

- ❑ Célibataire
- ❑ En couple

Si <> 2, ne pas poser 'CAR3: ACTIVITE DU CONJOINT'

Car2: Etes-vous actif ou inactif (chômeur de plus d'un an, retraité...)?

- ❑ Actif
- ❑ Inactif

Si ??CAR1: STATUT??=1 ET??CAR2: ACTIVITE DE L'INTERVIEWE??=1, spécifier la valeur '1' à la question 'Quota statut'
Si ??CAR1: STATUT??=1 ET??CAR2: ACTIVITE DE L'INTERVIEWE??=2, spécifier la valeur '2' à la question 'Quota statut'

Car3: Votre conjoint est-il actif ou inactif (chômeur de plus d'un an, retraité...)?

- ❑ Conjoint actif
- ❑ Conjoint inactif

Si ??CAR1: STATUT??=2 ET??CAR2: ACTIVITE DE L'INTERVIEWE??=2 ET??CAR3: ACTIVITE DU CONJOINT??=1, spécifier la valeur '4' à la question 'Quota statut'
Si ??CAR1: STATUT??=2 ET??CAR2: ACTIVITE DE L'INTERVIEWE??=1 ET??CAR3: ACTIVITE DU CONJOINT??=1, spécifier la valeur '3' à la question 'Quota statut'
Si ??CAR1: STATUT??=2 ET??CAR2: ACTIVITE DE L'INTERVIEWE??=1 ET??CAR3: ACTIVITE DU CONJOINT??=2, spécifier la valeur '4' à la question 'Quota statut'
Si ??CAR1: STATUT??=2 ET??CAR2: ACTIVITE DE L'INTERVIEWE??=2 ET??CAR3: ACTIVITE DU CONJOINT??=2, spécifier la valeur '5' à la question 'Quota statut'

Nous allons tout d'abord parler de vos achats de produits alimentaires, en commençant par les achats dans les grandes surfaces. Durant l'enquête, nous allons rechercher les magasins que vous fréquentez dans une liste qui comprend tous les commerces de la région. Cette recherche se fera avec votre aide et les outils de cartographie que j'ai devant moi. Plus vous serez précis pour m'indiquer un magasin, plus nous serons efficaces.

P1: Achats alimentaires GS

A1A. Pour vos achats alimentaires, dans quelle grande surface alimentaire (hypermarché, supermarché ou grandes surfaces spécialisées) allez-vous le plus souvent ?

1ère grande surface la plus souvent fréquentée

ENQUETEUR: UTILISER GEOCATI - SI NON REPONSE, SAISIR ZERO APRES UNE RELANCE

(doit être 0 entre 999999)

Si = , aller à 'Récup_nom'
Si >= , spécifier la valeur '??A1A. MAGASIN??[1]' à la question 'ORDRE_NEW'
Si ??A1A. MAGASIN??[1]=0, ne pas poser 'C2C. MODE TRANSPORT GSA'

A1A. Avec quelle fréquence allez-vous A1A. ENSEIGNES ?

ENQUETEUR: ENUMEREZ

- ❑ Plusieurs fois par semaine
- ❑ 1 fois par semaine
- ❑ 2 à 3 fois par mois
- ❑ 1 fois par mois
- ❑ Moins d'une fois par mois
- ❑ (NSP)

A1B. Quels produits avez-vous l'habitude d'acheter dans le magasin A1A. ENSEIGNES ?

ENQUETEUR: ENUMEREZ

(6 réponses maximum)

- ❑ Pain-Pâtisserie
- ❑ Fruits et légumes frais
- ❑ Charcuterie, viandes, volailles
- ❑ crustacés, poissons
- ❑ Surgelés
- ❑ Epicerie, crèmerie, autres produits alimentaires, produits d'entretien
- ❑ (Aucun de ceux-là)

A1A. Pour vos achats alimentaires, dans quelle grande surface alimentaire (hypermarché, supermarché ou grandes surfaces spécialisées) allez-vous le plus souvent ?

2ème grande surface la plus souvent fréquentée

ENQUETEUR: UTILISER GEOCATI - SI NON REPONSE, SAISIR ZERO APRES UNE RELANCE

(doit être 0 entre 999999)

Si = , aller à 'Récup_nom'
Si >= , spécifier la valeur '??A1A. MAGASIN??[1]' à la question 'ORDRE_NEW'
Si ??A1A. MAGASIN??[1]=0, ne pas poser 'C2C. MODE TRANSPORT GSA'

A1A. Avec quelle fréquence allez-vous A1A. ENSEIGNES ?

ENQUETEUR: ENUMEREZ

- ❑ Plusieurs fois par semaine
- ❑ 1 fois par semaine
- ❑ 2 à 3 fois par mois
- ❑ 1 fois par mois
- ❑ Moins d'une fois par mois
- ❑ (NSP)

A1B. Quels produits avez-vous l'habitude d'acheter dans le magasin A1A. ENSEIGNES ?

ENQUETEUR: ENUMEREZ

(6 réponses maximum)

- ❑ Pain-Pâtisserie
- ❑ Fruits et légumes frais
- ❑ Charcuterie, viandes, volailles
- ❑ crustacés, poissons
- ❑ Surgelés
- ❑ Epicerie, crèmerie, autres produits alimentaires, produits d'entretien
- ❑ (Aucun de ceux-là)

A1A. Pour vos achats alimentaires, dans quelle grande surface alimentaire (hypermarché, supermarché ou grandes surfaces spécialisées) allez-vous le plus souvent ?
3ème grande surface la plus souvent fréquentée

ENQUETEUR: UTILISER GEOCATI - SI NON REPONSE, SAISIR ZERO APRES UNE RELANCE

(doit être 0 entre 999999)

Si = , aller à 'Récup_nom'
Si >= , spécifier la valeur '??A1A. MAGASIN??[1]' à la question 'ORDRE_NEW'
Si ??A1A. MAGASIN??[1]=0, ne pas poser 'C2C. MODE TRANSPORT GSA'

A1A. Avec quelle fréquence allez-vous A1A. ENSEIGNES ?

ENQUETEUR: ENUMEREZ

- ❑ Plusieurs fois par semaine
- ❑ 1 fois par semaine
- ❑ 2 à 3 fois par mois
- ❑ 1 fois par mois
- ❑ Moins d'une fois par mois
- ❑ (NSP)

A1B. Quels produits avez-vous l'habitude d'acheter dans le magasin A1A. ENSEIGNES ?

ENQUETEUR: ENUMEREZ

(6 réponses maximum)

- ❑ Pain-Pâtisserie
- ❑ Fruits et légumes frais
- ❑ Charcuterie, viandes, volailles
- ❑ crustacés, poissons
- ❑ Surgelés
- ❑ Epicerie, crèmerie, autres produits alimentaires, produits d'entretien
- ❑ (Aucun de ceux-là)

ORDRE_NEW

NOM COMM

(stocke l'enseigne la plus fréquentée dans une variable pour la question D2A

Bibliography

[1] ABE Makato (1999), "Logistic Regression using the SAS System: Theory and Application," *Journal of Business and Economic Statistics*, Vol 17, 285-297.

[2] ALLISON, Paul D. (1999), "A Generalized Additive Model for Discrete Choice Data," SAS Institute Inc., Cary, NC, USA.

[3] ARNOLD, S. J., ROTH, V., TIGERT, D. J. (1981), "Conditional Logit versus MDC in the Prediction of Store Choice," *Advances in Consumer Research* Vol. 8, 1, 665-670.

[4] BAKER, Julie, PARASURAMAN, A., GREWAL, Dhruv, VOSS, Glenn B. (2002) "The Influence of Multiple Store Environment Cues on Perceived Merchandise Value and Patronage Intentions," *Journal of Marketing*, Vol. 66, No. 2. (April), 120-141.

[5] BASAR, Gozen, BHAT, Chandra (2004), "A parametrized consideration set model for airport choice: an application to the San Francisco Bay Area," *Transportation Research Part B,* 38(10), 889-904.

[6] BEN-AKIVA, Moshe, LERMAN, Steven R. (1985), "Discrete Choice Analysis: Theory and Applicaton to Travel Demand" Cambridge, MA: The MIT Press.

[7] BRIER, G.W. (1950), "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, Vol. 78, 1-3

[8] CLIQUET, Gérard (1995), "Implementing a subjective MCI model: An application to the furniture market," *European Journal of Operational Research*, Vol. 84, 279-291.

[9] CLIQUET, Gérard (2002), "Geomarketing: methods and strategies in spatail marketing" in "Geomarketing: methods and strategies in spatail marketing", G. Cliquet (Ed.), ISTE Publishing, Paris, France, 17-34.

[10] DION, Delphine, CLIQUET, Gérard (2002), "Consumer Spatial Behaviour" in "Geomarketing: methods and strategies in spatail marketing", G. Cliquet (Ed.) ISTE Publishing, Paris, France, 2002, 37-66.

[11] DOMENCICH, T., MCFADDEN, D.L. (1975), "Urban Travel Demand: A Behavioral Analysis", North-Holland Publishing Co.

[12] DUGUNDJI, E.R., WALKER, J.L. (2005), "Discrete Choice with Social and Spatial Network Interdependencies", *Transportation Research Record,* 1921, 70-78.

[13] FEATHER, Peter M. (1994), "Sampling and Aggregation Issues in Random Utility Model Estimation," *American Journal of Agricultural Economics*, Vol. 76, 772-780.

[14] FOTHERINGHAM, Stewart A. (1988), "Consumer Store Choice and Choice Set Definition," *Marketing Science*, Vol. 7, No. 3. (Summer1988), 299-310.

[15] GONZALES-BENITO, O. (2002) "Geodemographic and socioeconomic characterization of the retail attraction of leading hypermarket chains in Spain," International Revue of Retail, Distribution and Consumer Research," 12, 81Ű103.

[16] GUO, Jessica Y. (2004), "Addressing Spatial Complexities in Residential Choice Models," Ph.D Dissertation, The University of Texas at Austin.

[17] HAZIZA, David (2002), "Inférence en Présence d'Imputation Simple dans les Enquêtes: Un Survol," presented at the "Journées de Méthodologie Statistique," Paris.

[18] HECKMAN, J, "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models,", *The Annals of Economic and Social Measurement*, Vol. 5, 475-492.

[19] HENDEL, Igal (1999), "Estimating multiple-discrete choice models: An application to computerization returns," *Review of Economic Studies*, 66(2), 423-446.

[20] HOTELLING, Harold (1929), "Stability in Competition," *The Economic Journal*, Vol. 39, No. 153. (March), 41-57.

[21] HUFF, D.L. (1963), "A Probabilistic Analysis of Shopping Center Trade Areas," *Land Economics* 39, 81-90.

[22] KRIEGER, Abba M., POLLAK, Moshe, YAKIR, Benjamin (2003), "Surveillance of a simple linear regression" *Journal of the American Statistical Association*, Vol 98.

[23] LEBOUCHER, Séverine (2006), "Qui sont les champion ... du discount alimentaire?" *Le Journal du Management.*

[24] LITTLE, John D. C. (1970), "Models and Managers: The Concept of a Decision Calculus", *Management Science*, Vol. 16, No. 8, Application Series, (April), B466-B485.

[25] LE BLANC, David, LOLLIVIER, Stéfan, MARPSAT, Maryse, and VERGER, Daniel (2000), "L'Econometrie et l'étude des comportements: Présentation et mise

en oeuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT)," INSEE - Direction des Statistiques Démographiqes et Sociales.

[26] LESAGE, James P. (2000), "Bayesian Estimation of Limited Dependent variable Spatial Autoregressive Models", *Geographical Analysis*, 32(1), 19-35.

[27] LUCE, R.D. (1959), "Individual Choice Behaviour," New York: John Wiley & Sons.

[28] LUCE, R.D. (1977), "The Choice Axiom after 20 Years," *Journal of Mathematical Psychology*, 15, 215-233.

[29] MANSKI, Charles F. (1977), "The structure of random utility models," *Theory and Decision*, 8, 229-254.

[30] MANSKI, Charles F., MCFADDEN Daniel (1977), "Alternative Estimators and Sample Designs for Discrete Choice Analysis" in "Structural Analysis of Discrete Data with Econometric Applications", Manski, Charles, McFadden, Daniel (Ed.), The MIT Press, Cambridge, MA, 2-50.

[31] MARKLEY, Sébastien (2007), "Spatially-oriented discrete choice predictions: a Case Study of French supermarket preferences," *Case-Studies in Business, Industry, and government statistics*, 1(1), 26-46.

[32] MCFADDEN, Daniel (1974a), "Conditional Logit Analysis of Qualitative Choice Behavior," in "Frontiers in Econometrics," Zarembka, P., ed., New York: Academic Press.

[33] MCFADDEN, Daniel (1974b), "The Measurement of Urban Travel Demand," *Journal of Public Economics*, Vol 3., 303-328.

[34] MCFADDEN, Daniel (1978) "Modelling the Choice of Residential Location," in "Spatial Interaction Theory and Planning Models," Karlqvist, A., Lundqvist, L., Snickars, F., Weibull, J., North Holland: Amsterdam, 75-96.

[35] MONROE, K. B., GUILTINAN, J. P. (1975), "A Path-Analytic Exploration of Retail Patronage Influences," *Journal of Consumer Research*, 2, 2, 19-28.

[36] MURPHY, A. (1972), "Scalar and Vector Partitions of the Probability Score : Part I. Two-state Situation," *Journal of Applied Meteorology*, Vol. 11, 273-282.

[37] NAKANISHI, Masao, COOPER, Lee G. (1974), "Parameter Estimation for a Multiplicative Competitive Interaction Model: Least Squares Approach," *Journal of Marketing Research*, Vol. 11, No. 3. (August), 303-311.

[38] PUNJ, Girish N., STAELIN, Richard, *Journal of Marketing Research*, Vol. 15, No. 4. (November), 588-598.

[39] REILLY W.J. "The Law of Retail Gravitation", New York, W.J. Reilly, Inc.

[40] SEVERIN, V., LOUVIERE, Jordan, J., FINN, Adam, "The Stability of retail shopping choices over time and across countries", *Journal of Retailing,* (77), 185-202.

[41] SMITH, Howard (2004), "Supermarket Choice and Supermarket Competition in Market Equilibrium," *The Review of Economic Studies,* 71, 253-263.

[42] SMITH, Howard (2006), "Store Characteristics in Retail Oligopoloy," *RAND Journal of Economics*, Vol. 37, No. 2 (Summer), 416-430.

[43] SMITH, Tony E., and LESAGE, James P. (2004) "A Bayesian Probit Model with Spatial Dependencies," in "Advances in Econometrics: Volume 18: Spatial and Spa-

tiotemporal Econometrics," James P. LeSage and R. Kelley Pace (eds.), Oxford: Elsevier Ltd, 127-160.

[44] SWAIT, J., BEN-AKIVA, Moshe (1987), "Incorporating random constraints in discrete models of choice set generation," *Transportation Research B,* 21(2), 91-102.

[45] THOMAS, Alban (2000), "Econométrie des Variables Qualitatives," *DUNOD*, Paris.

[46] TRAIN, Kenneth (2003), "Discrete Choice Methods with Simulation," Cambridge University Press.

[47] TRAIN Kenneth E., MCFADDEN Daniel L., and BEN-AKIVA Moshe (1987), "The Demand for Local Telephone Service: A Fully Discrete Model of Residential Calling Patterns and Service Choices," *The RAND Journal of Economics*, 18, 109-123.

[48] TUROLLA, Stéphane (2007), "Compétition Spatiale dans le Secteur de la Grande et Moyenne Distribution Française," document de travail, LASER, Université Montpellier I.

[49] WALKER, Joan L., BEN-AKIVA, Moshe, BOLDUC, Denis (2007), "Identification of Parameters in Normal Error Compônent Logit Mixture (NECLM) Models," *Journal of Applied Econometrics*, 22(6), 1095-1125.

Abstract

The thesis was done in collaboration with the BVA Institute, a survey company that hoped to develop techniques of forecasting French spending based on data sets from their own consumer surveys. We developed a Conditional Logit model in order to predict the large surface stores chosen by each household, and used imputation in order to predict the products they chose. Since store choice was insensitive to household caracteristics, the use of home-store distances and the geographic characteristics of store neighborhoods was essential to our predictions.

In the first chapter, we present Logit Models in general, and describe the data that we use to apply our modeling techniques.

In the second chapter, we explore how we adapt the Conditional Logit model to choices of stores. Due to the fact that a choice of store has too many alternatives for estimation to be tractable, we test several modifications of our model that either reduce the size of each choice set, or that result from random draws of the alternatives. Since traditional evaluation methods based on likelihood were inappropriate for comparing these different techniques, we developed a criteria based on the model calibration to choose the best estimation technique.

In the third chapter, we present the results of our estimations on our sample, presenting the technique that shows the best trade-off between predictive accuracy and cost of use.

In the last chapter, we look at the use of imputation in order to predict product choice based on store choice.

Keywords: Discrete choice, Supermarkets, Spatial data

Abstrait

Cette thèse CIFRE a été réalisée au sein de l'institut de sondage BVA. BVA développe des techniques de prédiction de la répartition des dépenses françaises à partir de bases de données de consommation. Dans ce cadre, nous avons construit un modèle Logit Conditionnel pour prédire les choix de magasins de grandes surfaces des ménages, puis utilisé les techniques d'imputation pour prédire les choix de produits de ces mêmes ménages. Nous montrons que les choix de magasins sont insensibles aux caractéristiques sociodémographiques des ménages. Par contre, l'utilisation des distances entre magasins et domiciles et les caractéristiques géographiques des voisinages des magasins sont essentielles pour la prédiction.

Dans un premier chapitre, nous rappelons les principaux aspects des modèles Logit Conditionnels, et décrivons les données utilisées.

Dans un deuxième chapitre, nous adaptons le Logit Conditionnel au problème traité. Nous explorons différentes pistes pour réduire la taille trop importante de l'ensemble de choix. Puis, après avoir étudié les propriétés des critères usuels d'évaluation de la prédiction dans les modèles de choix, nous proposons un autre critère basé sur la calibration du modèle.

Dans un troisième chapitre, nous donnons une illustration à partir des données de l'enquête " Flux d'Achats " sur la Région Centre.

Dans un dernier chapitre, nous utilisons les techniques d'imputation pour prédire les choix de produits selon les choix de magasins.

Mots clés: Choix discrèts, Supermarchés, Données spatiales