CS-BIGS 1(1): 26-46 http://www.bentley.edu/csbigs/vol1-1/markley.pdf

Spatially-Oriented Discrete Choice Predictions: A Case Study of French Supermarket Preferences

Sébastien Markley
GREMAQ, Université de Toulouse I, France

Introduction

Given the importance of the economic stakes involved in understanding and predicting the behaviour of grocery consumers, it is not surprising that economic papers and research projects involving the modelling of decisions related to the purchase of food products abound (Smith, 2004; Erdem, Imai, and Keane, 2003; Hendel and Nevo, 2006). One common technique used, described in Ben Akiva and Lerman (1985) and Train (2003) is discrete choice modelling, in which a consumer's choice is precisely defined (for example: choice of store franchise within one type of store, choice of products within a category of product, choice of brand for one product, etc) and then a statistical methodology is devised to predict the choice made by the consumer in a manner consistent with previously observed shopping behaviour.

Within the realm of discrete choice models of grocery consumption, the choice of store has the added property of being spatially-oriented, since it depends on the distribution of home and store locations. In this paper, we will present a model of the choice of large-surface stores by households in France as it arises in a real-world setting. Since the data used include geographic co-ordinates of stores and households, our approach will involve a study of the effect of geography on behavior. We also present a data set based on a survey of household consumption on which readers can practice making use of these techniques themselves, and we explain the statistical programs used to arrive at our results.

This paper originates from work done for the French survey institute BVA in the development of statistical tools for the prediction of market demand. In an earlier project, BVA produced a model to predict the probability that a given individual would purchase products within a given class of products (for example, gardening implements) and then predict the amount of money that these individuals would be expected to spend on these purchases, if they made them. Once the model was developed, it was used to predict the choices made by every individual in a data set provided by INSEE of a sample of one-twentieth of individuals enumerated in the French census of 1999. With this information, for just about any geographic subdivision of France of a reasonable size (usually with a population of more than 2000 individuals), BVA was able to produce an estimate of the number of individuals purchasing the given products, and how much money would be spent on them. This was a powerful tool for predicting total market demand, but it was apparent that what was also vital to the retailer was the question of how the purchases of goods by a given set of consumers were distributed amongst the suppliers. As an extension of this project, BVA sought a model that could predict the grocery store in which a given household chose to make its purchases, and what factors would go into this decision.
[2]

To develop such a model, we make use of a survey that BVA undertook in which each household in a chosen sample was asked which large-surface store, or store having at least 300 m^2 of retail space it visited most often for food purchases, which large-surface store it visited the second-most often, and which it visited the third-most often. In cases where households visited less than three large-surface stores, it could answer "no store" for the appropriate choice of store. The basis of our model of large-surface store choice is the prediction of the answers to this set of questions. Our hope is that this model can be used to predict the expected clientele of any large-surface grocery store in France with the use of publicly available census data on the surrounding population and with the use of data on the characteristics and geographic distribution of French supermarkets.

We shall present the application of this methodology to data from a set of households in the Indre-et-Loire Department in France. In Section I, we describe the modeling techniques that we use. This involves discussing Conditional Logit models in general, and a discussion of the problem of defining choice set size. In Section II, we introduce the data set that is available to the reader, and to which our model is applied. We define the geographic variables that we use in our model, and use maps to show how the values of these variables are distributed amongst the communes and survey sectors of the department. We also show the locations and concentrations of large-surface store types in the department. In Section III, we show how our model is calculated using the SAS program, and which explanatory variables we specify in our model. We discuss how model estimations can be interpreted and evaluated, and then use cluster analysis to present the model's predictions. The last section of the paper is devoted to discussing measures of the reliability of the model's predictions.

I. Method

Our strategy is to break down the households' choices of stores into three separate decisions: the choice of the store most-visited by the given household, the store the second-most visited, and the store the third-most visited. What we are predicting is one of the following probabilities for all individuals i and stores j:

P1(i,j) = P(individual i will select the large-surface store j as the large-surface store it visits the most often for grocery purchases.)

P2(i,j) = P(individual i will select the large-surface store j as the large-surface store it visits the second-most often for grocery purchases.)

P3(i,j) = P(individual i will select the large-surface store j as the large-surface store it visits the third-most often for grocery purchases.)

We model these three decisions independently. Clearly, supposing P1, P2 and P3 to be independent is unrealistic. There are obvious technical reasons for this; for one thing, we do not allow households to name the same large-surface store for more than one of the three choices of stores. Also, households selecting no store for the first and second choice of store necessarily select no store for the subsequent choices of stores. However, there are more fundamental reasons for non-independence. We find that when households choose more than one large-surface store, they tend to choose different store types whose dissimilar trade-offs in terms of accessibility and convenience will allow a more flexible adaptation of their shopping trips to their schedules. For example, it is more logical for a consumer to visit a large hypermarket and a supermarket, than two supermarkets, since this consumer can use a vehicle for occasional trips to buy non-perishable goods in the hypermarket while going on foot several times a week to buy perishable items from the supermarket. We find that the result of an independence assumption is an overestimate of the number of households selecting two or more large-surface stores of the same type, and an underestimate of the number of households selecting more than one store type. For this reason, we have considered some ways of introducing the dependencies of the large-surface store choices into the model.

We thought of creating triplets of stores for each household, containing the most-visited, second-most visited, and third-most visited stores, and using one conditional logit model to predict the probability that each household behaves according to each triplet. However, the number of different combinations of store choices is too great to make this feasible and likely correlations between error terms within alternatives in the same choice set pose a problem. A more promising option would be to construct explanatory variables in the prediction of one of the choices of large-surface store from the model estimations of the other choices of large-surface stores. For example, we could create a variable for each household that records the store type of the large-surface store with the highest predicted probability of being selected as the first choice of store most-visited. This will then be entered as an explanatory variable for the model of the probability of selecting the store that is the second-most visited. Another idea is to take the probabilities of selection of the first choice of store to define subpopulations of our sample on which the models of the second choice of large-surface store can be run independently. The best approach we found was to model the second and third choices of large-surface store conditioned on the store type of the first choice of large-surface store. All these techniques greatly complicate our model, thus rendering the estimated parameters of our model less clear to interpret and rendering the process of selecting explanatory variables and adapting the model to the data more time-consuming. We have therefore dropped the consideration of dependence between large-surface store choices from our paper for ease of presentation.

A. Conditional Logit modeling

For each of the three choices of large-surface store, we make use of a Conditional Logit model to predict the households' decisions. For a detailed explanation of discrete choice modeling, the reader is advised to look at Ben-Akiva and Lerman (1985) and Train (2003). We begin by assuming that an individual, when faced with the choice of one of several mutually exclusive alternatives, assigns an imaginary value called a utility (which can be thought of as attractiveness) to each choice and then chooses the alternative with the largest utility. This means that if the utilities of a set of alternatives can be determined, then the choice of the individual can be predicted. Where the utilities cannot be determined, we can assume that they follow a random distribution. Once we do this, even if we can't predict the decision made by the individual, the probability that a given alternative will be selected will be the probability that the utility of the alternative is greater than the utilities of all other alternatives in the same choice set. A model of the probability of selection of an alternative in a discrete choice set that is based on randomly distributed utility terms is called a random utility model.

A Conditional Logit model is a random utility model in which we assume that the utility is the sum of two independent components: the systematic component, which is a deterministic function of known variables, and the disturbance term, which is randomly distributed. Schematically, if U_{ij} is the utility of alternative j for individual i, V_{ij} is the systematic component, and ε_{ij} is the disturbance (or error) term, then

$$U_{ij} = V_{ij} + \varepsilon_{ij} \tag{1}$$

If V_{ij} is a linear combination of variables representing what is known about the alternative j for individual i (the explanatory variables) we can represent this with the equation

$$V_{ij} = X_{ij}\beta \tag{2}$$

X_{ij} is a vector containing the values of the explanatory variables. It is multiplied with the vector β, which represents the coefficients of the terms in X_{ij}. According to the Conditional Logit model, the terms of the vector $\varepsilon_{ij,}$ are independent and identically distributed (iid) and follow the extreme-value distribution. The validity of the iid assumption depends upon our ability to choose a set of variables that we include in X_{ij} that account for the factors involved in the choice of alternatives made by the individual without introducing irrelevant information. The extreme-value distribution is the limit distribution of the maxima of a series of independent and identically distributed random variables and works as an approximation of the normal distribution that is vastly superior in terms of mathematical simplicity and ease of calculation. The maximum of a series of extreme-value distributed random values is also extreme-value distributed, a property that allows for a simple, closed-form calculation of the predicted probabilities of selection in our discrete choice model. The probability that the individual *i* chooses alternative *j* (out of a set of possible alternatives J_i for the alternative *i*) will be:

$$\begin{aligned} P(i\ chooses\ j) &= P(U_{ij} \geq U_{ik}, \forall k \in J_i) \\ &= \frac{\exp(X_{ij}\beta)}{\sum_{k \in J_i} \exp(X_{ik}\beta)} \end{aligned} \tag{3}$$

However, before we can use this equation to predict the choices of individuals, we need to determine the values of the coefficient vector β. This vector will always remain unknown, but if we have a sample of individuals for whom both the relevant explanatory variables for each alternative presented to it and the actual choice of alternative made are known, then we can use maximum likelihood estimation to choose the values of the β vector that predict the probabilities of selection for the individuals in the sample that best correspond to the behaviour observed.

B. Defining the choice set

We have already seen that in a Conditional Logit model, the choice set, or the term J_i, must be known for every individual i and all individuals must select one and only one alternative in the choice set. That means that if we are dealing with a choice of large-surface stores, then for our model to be coherent, we must include in J_i enough large-surface stores that it would be impossible for the household's choice of store not to be contained in J_i. The problem is that for any household, there are a very large number of large-surface stores from which it is possible to choose. According to the model equation, we have to enter a vector of explanatory variables for every single alternative in the choice set, so the amount of data required for large choice sets can quickly render model estimation intractable. What we have done in response to this problem is to redefine our choice set to include a limited set of alternatives representing the large-surface stores we know the individual is likely to choose, and the alternative representing all other choices of large-surface stores. Care must be taken with the definition of our choice set, as it involves trading off valuable properties of our model. Larger choice set sizes create more informative predictions (as we have more predictions of selections of individual stores) yet also involve much greater computational burden, and not necessarily greater predictive accuracy, if the additional alternatives added to the choice set have little likelihood of being selected by the individual. The smaller the choice set, the more likely the household will be entered as selecting the "outside" option for which no information on the choice selected is recorded. Howard Smith (2004) used as his initial choice set the 30 large-surface stores that were closest to the household's home, and included a 31st alternative representing an outside option. However, by taking into account the store type in the creation of our choice set, we could decrease the size of the choice set, while also decreasing the likelihood that the household would select the outside option. We have examined different sizes of choice sets in previous work, and decided upon a choice set of about 12 alternatives as being the best (Markley, 2006b).

II. Data

A. Source of data: a survey of shopping behavior

In the spring of 2004, the survey institute BVA undertook a survey of shopping behavior in the Centre Region of France. A total of 14,217 households were selected for whom a detailed questionnaire was filled out. Once our sample of households was chosen, interviews were sought with individuals within the households in order to fill out a questionnaire detailing their shopping behavior, and asking which three large-surface stores the household visited most often for food purchases. The survey also provided detailed information on the household's characteristics, including socio-professional category, age, and access to transportation. In addition to this, we had access to a data set containing the large-surface grocery stores within the Centre Region and also in all the French departments bordering the Centre Region.

There are three geographic units used in our sample that we take the time here to explain: survey sectors, communes, and IRIS. Survey sectors are the smallest units at which the survey sample is representative. Communes are the smallest units for which many of the variables used in the survey are defined. And the IRIS are the geographic units used for the assignment of geographic co-ordinates to households' homes.

The survey selection was done following stratified quota sampling. The survey area was divided geographically into survey sectors chosen so as to contain roughly the same number of households, within one order of magnitude, and be homogeneous in terms of behavioural characteristics, and in terms of the large-surface stores that were accessible. In the Centre Region, the populations of each survey sector range in size from 1080 to 10100 households, with three-quarters of sector populations containing between 2200 and 4100 households. Figure 1 shows the distribution of the populations of the 56 sectors in the Indre-et-Loire department.

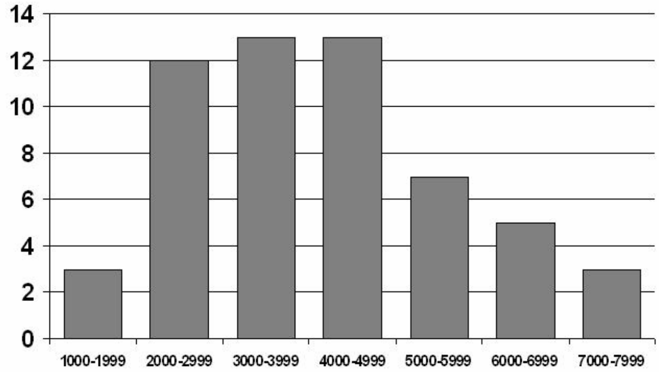

Figure 1 : Histogram of sector populations in the Indre-et-Loire department

The smallest class of administrative district in France is the commune. Communes correspond to municipal governments and so a lot of French statistical data, including some of the variables used in our model, are calculated at this level. Most communes are rural, with small populations, but some represent large cities and must be divided into smaller geographic zones for the purposes of data collection. This is why we use a finer geographic zone than the commune called the IRIS.

With every census, INSEE, in co-operation with the governments of each commune, divides the French territory into "Ilots" or geographic units determined by the features of the land at the time. These "Ilots" are then aggregated to form continuous geographic zones for the census in question. These zones are called IRIS. We represent the distribution of the 1624 IRIS represented by our survey sample in Figure 2. We see that the points on our scatter plot are divided into two clusters, one representing IRIS with small populations and a relatively large size, grouped to the lower left of the scatter plot, and another representing very small IRIS with large populations, clustered along the X axis. The latter represent urban IRIS and are generally small, densely populated sectors of inner cities.

Because the primary factor in determining a choice of store was the distance a household needed to travel to a store, the most important data that we collected in our survey were the geographic co-ordinates of the household's home and the stores listed in our survey area, which could enable us to create a data set that contained the Euclidean distance between each home and each store, and more importantly, determine which large-surface store was the closest to each household.

This kind of information is in general very expensive, and is not often available to those studying shopping behavior, so we were very pleased to make use of it in this study. Unfortunately, although we recorded the addresses of the households interviewed in our survey, the cost of transforming addresses into exact geographic co-ordinates was far too expensive to be done. We therefore took as the co-ordinates of the household's home the centroid of its IRIS of residence, or the center of mass of the population of the IRIS in cases where this corresponded to a single commune. This obviously meant that many households were assigned the exact same geographic co-ordinates.

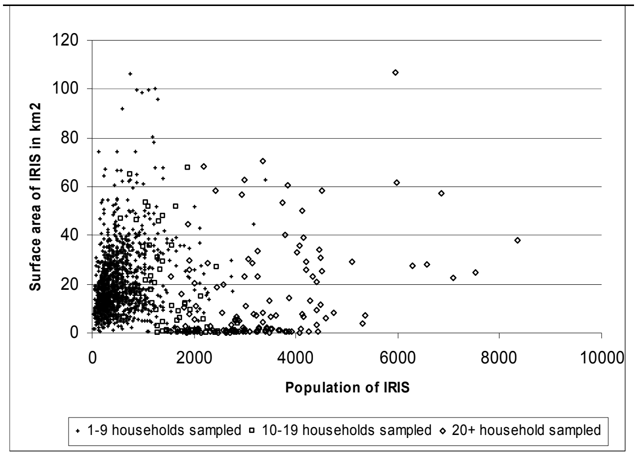

Figure 2 : Distribution of IRIS represented by our sample by surface area, population, and number of households selected for the survey sample.

The geographic co-ordinates of the stores in our survey, on the other hand, were more precise, corresponding to the centroid of a polygon drawn around the commercial zone in which the store was located. Neighboring stores were usually assigned the same geographic co-ordinates.

The imprecision of the co-ordinates of the households could be a source of model error, especially if it were great enough to cause us to be mistaken in the determination of the store closest to the household. We must also bear in mind that the Euclidean distance between a household and a large-surface store does not translate exactly into travel times between homes and stores. 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.

B. The Indre-et-Loire department

The data set that we use in this paper, and that the reader can work with, is a small subset of the data from our survey. This data set contains the households living in the Indre-et-Loire department of France (Department 37) located a little less than 200 km southwest of Paris and including the city of Tours on the Loire River. We need to ensure that the model presented in this paper resembles the model that was developed for BVA on the entire data set, in terms of its estimated parameters and its predictions. Because estimated standard errors tend to be greater when calculated over smaller samples, it is necessary to eliminate many variables from the model run on the Indre-et-Loire department that are found in the model run on data from the entire region, in order to ensure that all the effects included in our model are significant. We also eliminate terms reflecting interactions between main effects in order to increase the interpretability of our parameter estimates.

In Figure 3, we include a road map of the Indre-et-Loire department showing us the main transportation axes. The department contains only one important urban centre in Tours. This city is on the Paris-Bordeaux freeway that links Tours with the city of Blois to the northeast and the city of Châtellerault to the south.

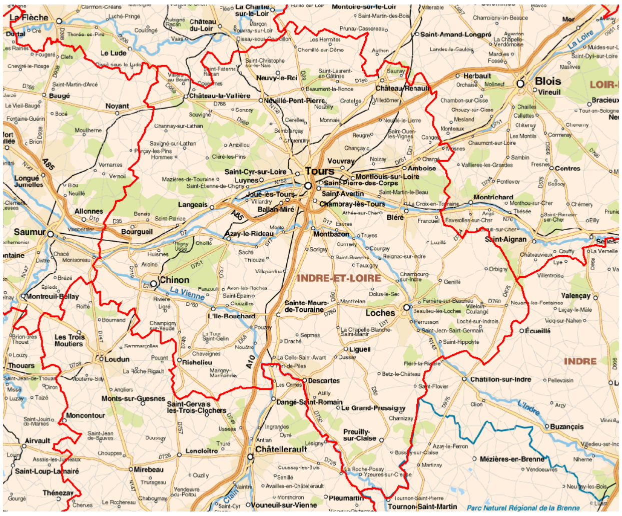

Figure 3 : Road map of the Tours area with the Indre-et-Loire Department outlined.

C. Types of large-surface stores

Large-surface food stores in France are generally divided into three types: 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. 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. There are four large hypermarkets in the Indre-et-Loire department, all of them in the Tours urban area.

Hard discount stores are identified as belonging to a brand that follows the hard discount business model. They are still relatively new in France, but are rapidly expanding their market share and changing the dynamics of food retailing in France. 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 travel burden. In just five years, from 2000 until 2005, the market share of hard discount stores in food purchases in France has gone from 9 to 13.3 percent. At the same time, the percent of French households visiting hard discount scores went from 55.3 percent to 66.8 percent between 2000 and 2004 (Leboucher, 2006).

Table 1. Number of stores of each type in the Indre-et-Loire department

Store Type	Number
Supermarket	75
Large Hypermarket	4
Hard Discount	26
Small Hypermarket	10

D. Exploratory statistics and choice set definition

Our sample contains 3968 individual households in the Indre-et-Loire department, each choosing one of the four types of large-surface stores (or none). We have used the following coding for these categories: supermarkets (SM), small hypermarkets (HM), hard discount stores (HD), and large hypermarkets (XM).

The following charts were calculated from the responses recorded for our Indre-et-Loire sample. In Figure 4, we see that the order of the store choice has an important role in determining the type of store the household chooses. We note that all but 2 percent of households chose at least one large-surface store for its shopping needs, 75 percent of the population chose two or more and only 31 percent chose three stores. The charts show the percent of households choosing each type of large-surface store for the first, second, and third choice of large-surface store conditional on there being a store visit.

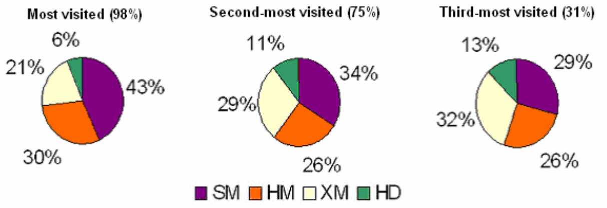

Figure 4 : Proportion of households selecting each type of large-surface store among those visiting a large-surface store for each order of store choice.

We would expect that households would tend to minimize the cost of a shopping trip, so we would expect a household to choose the closest store to its residence, all other factors being equal. And indeed, we find that 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 store to the household's co-ordinates (the centroid of the household's IRIS of residence). In Table 2, we look at the choices of stores within each category of large-surface store, and we calculate what percentage of each category corresponds 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 46 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 co-ordinates. However, only three 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.

Table 2. Breakdown of choices of each type of large-surface store by rank of distance of store from households' home.

	SM	HM	HD	XM	All
Closest Store	46%	8%	13%	3%	22%
2nd Closest	17%	12%	10%	4%	11%
3rd Closest	7%	8%	7%	5%	7%
Other Stores	30%	72%	70%	88%	60%

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 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 clearer. In Table 3 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 55 percent of the time a household chooses a supermarket as 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 behavior pattern seems to be less well-maintained for hard discount stores.

Table 3. Breakdown of choices of each type of large-surface store by rank of distance of store from households' home calculated with respect only to other large-surface stores within the same category of large-surface store.

	SM	HM	HD	XM	All
Closest Store within store type	55%	57%	36%	60%	55%
2nd Closest within store type	16%	28%	19%	20%	21%
3rd Closest within store type	6%	5%	9%	12%	8%
Other Stores within store type	23%	10%	36%	8%	26%

We mentioned above that the geographic co-ordinates of the households' homes and large-surface stores we use are imprecise. 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 evenly 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 that, if assigned the true geographic co-ordinates of their homes, would have the closest and second-closest large-surface stores in a different order than with the current, less-accurate co-ordinates. 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. The results of these calculations are in Table 4. We can assume that rendering our geographic co-ordinates more accurate would have almost no effect on the correct determination of the closest hypermarkets to the household's home, although this could have an effect on supermarkets.

Table 4. Expected percent of households in nonrural IRIS for whom the rank of the distances of large-surface stores does not change with the replacement of the assigned geographic co-ordinates of each household by the true co-ordinates.

	Well-ordered	Std Dev
Supermarket	86%	0.3%
Hypermarket	95%	0.2%
Hard Discount	89%	0.2%
Small Hypermarket	97%	0.1%

Besides showing us that people frequently shop in nearby stores, these tables show us that for most stores in a given household's choice set, the probability of selection is extremely small and little will be gained by having a model that attempted to predict it accurately. We have therefore reduced the size of our choice set by aggregating all stores for whom we believe a priori that the probability of selection will be small into a category labeled "other stores". The large-surface stores that are not aggregated are the stores in each category of store that are closest to the households' home.

We have decided which supermarkets to include in the "other" option by taking the distances of all the supermarkets in a choice set from the household's home, and ranking them from closest to furthest. The ranks were calculated by adding one to the number of large-surface stores that were closer to the household's domiciles. We decided, based on the percentages listed in Table 3 that we would include supermarkets and hard discount stores that had a distance rank of three or less, and small hypermarkets and large hypermarkets were included if they had a distance rank of two or less. All other stores were included in the category of other stores. It is important to note that this choice set would typically include 12 options, but could include more than 12, due to tied distances.

We have then calculated the percentages of households in our sample selecting each category of alternatives for each choice. The sums of the percentages for each choice add up to 100. In Figures 5 and 6, SM1 refers to supermarkets of distance rank 1, XM2, large hypermarkets of rank 2, etc. OUT refers to "other stores", and NO refers to the choice of choosing no store. Obviously since there are far more non-choices in the third than the first and second choices, the percent of households choosing a supermarket as its second choice will be lower than the percent of households choosing a supermarket for its first choice.

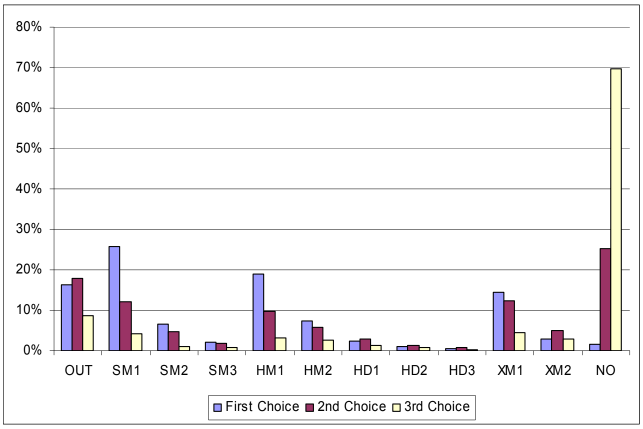

Figure 5 : Frequencies of alternative selections conditional on order of choice

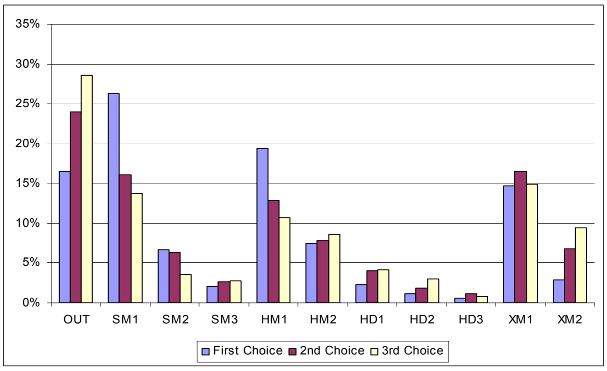

Figure 6 : Empirical Probabilities of alternative selections conditional on order of choice and on selection of a large-surface store

We have also recalculated the percentages conditional on the household selecting a large-surface store. This gave us the graph in Figure 6. We see here that unsurprisingly, the probability of choosing a store that is the closest in its category drops greatly as we go from the first to the second and third choice of stores. This could be explained partly by the fact that if a household chose the closest store as its first choice, it cannot choose the same store as its second choice. We also see that the probabilities of choosing the OUT option increase as households are more likely to go further for their subsequent choices of stores.

III. Model estimation and predictions

A. Variables

We dispose of a great deal of information that we could use in order to predict the choice of large-surface store by individual. We can divide the variables we have at our disposal into three categories: 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.

We believe that there are three basic factors determining a household's choice: the taste preferences of the individual, the attractiveness of the alternatives available to the household, and the cost of accessing each alternative. We have not found socio-demographic variables characterizing individuals to be very useful in predicting store choice (Markley, 2006b). This could be because these variables do not adequately capture taste variation since an individual's taste is a psychological phenomenon that is too complex and individualistic to be reduced to broad socio-demographic factors. We do find that the store's attractiveness, in terms of the store's general type, the store's size in retail space, and the store's name play roles in determining the store choice. However, by far the most important factor, once the type of store is chosen, is the accessibility of the store. This accessibility is not only represented by the distance of the store from the home (in both rank of distance, and absolute distance) but also by a variety of factors that represent the geographic and economic links between the communities in which stores are located. 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 that household members tend to go to for work, study, or leisure. We do not have direct information on where household members work, go to school, or spend their evenings, but our geographic variables can help identify the areas that are more likely to attract household members for these purposes.

It is important to note that all the explanatory variables that are chosen for our model in this paper are invariant within the IRIS of the household in question. This means that all households coming from the same IRIS will have the same predicted probabilities of selecting each choice. This was not intended, but it is a consequence of the elimination of socio-demographic variables from our model. This would justify creating a model done at the level of survey area, rather than at the individual level, but BVA requires the assignment of stores to individuals, for the purpose of providing a basis for the assignment of other behavior patterns (e.g. frequencies of store visits and products bought) to these individuals later.

B. Model equation

Of all the available variables, we select eight that not only serve well to predict the store choice, but produce coefficients that can be interpreted intuitively. These effects include the following (seen in Table 5):

- Store type
- Rank of store distance, given store type
- Euclidean distance between household's domicile and store
- Retail space of store in thousands of square meters
- Polarity of commune in which the store is located
- Commune of store is preferred destination of households in commune of households' home
- Population density of commune in which store is located
- Store is in same department as household's home
- Store is in same commune as household's home

Polarity refers to the INSEE classification of communes into four classes. An urban pole is a commune that provides employment to residents of surrounding communes. Monopolarized communes are communes not in urban poles but whose residents tend to work in one urban pole. Multipolarized communes are communes not in urban poles and not monopolarized whose residents tend to work in several urban poles.

Table 5. Glossary of explanatory variables used in the discrete choice model.

Variable	Definition	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
Nostore	No store	Dich
disSM	Euclidean distance of supermarket from home in meters	Cont
disHM	Euclidean distance of small hypermarket from home in meters	Cont
disHD	Euclidean distance of hard discount from home in meters	Cont
disXM	Euclidean distance of large hypermarket from home in meters	Cont
surfSM	Surface area of supermarket in thousands of m^2	Cont
surfHM	Surface area of small hypermarket in thousands of m^2	Cont
surfHD	Surface area of hard discount in thousands of m^2	Cont
surfXM	Surface area of large hypermarket in thousands of m^2	Cont
gsVC99_1	Commune of large-surface store classed as city centre.	Dich
gspol99_1	Commune of large-surface store classed as urban pole.	Dich
gspol99_12	Commune of large-surface store classed as urban pole or monopolarized.	Dich
gspol99_123	Commune of large-surface store classed as urban pole, monopolarized, or multipolarized.	Dich
gspol99_23	Commune of large-surface store classed as monopolarized, or multipolarized.	Dich
gspol99_4	Commune of large-surface store classed as rural.	Dich
FavCom	Commune of large-surface store is preferred destination for residents of household's home commune.	Dich
Denspopu	Population density of commune of large-surface store	Cont
Samedep	Large-surface store is in same department as household's residence	Dich
Samecit	Large-surface store is in same commune as household's residence	Dich

Finally, nonpolarized communes are communes whose residents don't tend to work in any urban poles.

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 last column indicates whether the variables are dichotomous (Dich) or continuous (Cont). We note that the first 12 variables "SM" through "nostore" are choice-specific constants identifying which alternative is being selected by the household. Their coefficients provide an estimate of the unaccounted attraction of the store type and the rank of the store distance. The variables SM, HM, HD, XM, outside, and nostore are linearly dependent variables. The variables gspol99_1, gspol99_12, gspol99_123 are indicator variables for the variable gspol99 indicating within which of the four categories of polarity the commune in question belongs.

The variable disSM is zero when the store in question is not a supermarket. The other variables beginning with dis and surf are defined in the same way.

C. SAS program

The following is the SAS program used to calculate the maximum likelihood estimates of our model from the recorded choices of large surface stores in our sample of households in the Indre-et-Loire data set. We call this sample the training data set, since it is used to "train" our model to do predictions that are based on the shopping behaviour it represents.

```
proc mdc data = DataFile;
      title "Model of first choice of
store for households in Indre-et-Loire";
      model lieuchx1 =
            SM SMRankGE2 SMRankGE3
            HMRankGE2
            HD HDRankGE2 HDRankGE3
            XM XMRankGE2
            outside nostore

            disSM disHD disHM disXM
            surfSM surfHD surfHM surfXM

            gsVC99_1 gspol99_1 gspol99_23
            FavCom denspopu
            / type = clogit choice =
(AltID) maxiter = 400;
      id cle1;
      ods output Parameterestimates =
ParmEsts1;
      output out = PredProbs1 (keep = cle1
Lieu AltConst Pk1 Lieuchx1)
p = Pk1 xbeta = xb;
run;
```

This procedure is called MDC, after "Model of Discrete Choice", and it has been designed to perform calculations of a variety of different types of multivariate random utility models including conditional logit, nested logit, HEV (Heteroskedastic Extreme Value), and multinomial probit models. It is the procedure best suited to doing conditional logit models. (This procedure is only available in Version 9 of SAS. In previous versions of SAS, the same calculations could be done using the PHReg Procedure. The syntax is only slightly different, and the variable indicating the chosen option takes a value of 1 for the alternative that is chosen by the individual, and 2 for the other alternatives.) We shall go through the details of this program one step at a time.

- The reader will notice that the first line of the program includes the statement "data = DataFile" that specifies the name of the file attached to this paper. This statement specifies the file containing the data on which the maximum likelihood estimates will be calculated.
- The file must contain one observation for every alternative available to every individual in the data set. Each individual making a decision (in our case households) is identified with a unique value of the identifier variable. The unique identifier variable must appear in the "id" statement that we see in the fourth line from the bottom of the program.
- In the DataFile file, each household is identified with a value of the variable cle1.
- For every single decision-maker, the alternatives must each be identified with another variable that appears in the statement choice "= (AltID)" in the fifth line from the end of the program. The program cannot function if one decision-maker has two alternatives in his/her choice set with the same value of this variable.
- The variable that identifies our choices of alternatives is called AltID. The first digit of this five-digit variable contains the identification of the store type:

  1 = Outside store
  4 = Supermarket
  5 = Small Hypermarket
  6 = Hard Discount
  7 = Large Hypermarket

9 = No store

The next two digits identify the rank of the distance of the store amongst stores of the same type. The last two digits are an arbitrary designation designed to differentiate between two stores of the same type and of the same rank.

- The lieuchx1 variable indicates the choice that is observed for the given observation. This variable is one for the choice observed, and zero for any other choice. The values of this variable must sum to one for every single household in the data set.
- In the model statement, only one variable may be specified on the left-hand side of the equality. On the right-hand side of the equality is a list of the explanatory variables that determine the non-random components of utility. Here, we can see the list of variables that we chose.
- It is very important that we not have colinearity or near-colinearity between the explanatory variables, since this will cause problems in the model calculations. For this reason, we have removed the variables HM and gspol99_4, as they are both linear combinations of other variables in the list.
- At the end of the model statement, "type = clogit" specifies that the type of random utility model we are using is the Conditional Logit model where explanatory variables describe the characteristics of the alternatives, rather than the individual decision-makers, and where choice sets can vary by individual.
- The procedure uses a Newton-Raphson algorithm in order to calculate the maximum likelihood estimates of the model coefficients, and the statement tells the "maxiter = 400" program to halt the calculation after 400 iterations if it does not converge. If the program does not converge, then the program issues a warning in its output and the user must note that the estimated coefficients are not maximum likelihood coefficients.
- The last two lines of our program give us the names of the output data sets created by our program. The statement "ods output Parameterestimates = ParmEsts1" tells the program to store the values of the estimated parameters of the model in a SAS dataset with the name ParmEsts1 along with the variance and p-values of the estimates. The statement "output out = PredProbs1 p = Pk1" tells the program to create a SAS dataset in which all the values of the DataFile data set are copied, but to which a new variable, named Pk1 is added, which contains the values of the probabilities of selection of each alternative for each individual calculated using our model equations with the maximum likelihood estimates of our parameters as the model's parameters.

D. Results

When we run the above SAS program on our training data set, we usually get two or three pages of output. This information is under the headings "Model Fit Summary", "Discrete Response Profile", "Goodness-of-Fit Measures", and "Parameter Estimates". What we check first of all is if we find the statement "algorithm converged" at the beginning of the model printouts. We can also check if "Number of observations" or the number of distinct values of the variable cle1 corresponds to the number of individuals in our sample, and the "number of cases" corresponds to the number of observations in our file since SAS eliminates observations where some of the explanatory variables have blank values.

The "goodness-of-fit" measures are all measures of the degree to which the predictions of the probabilities of selecting each alternative for each individual in our data set reflect the choices observed for each individual. The SAS output helpfully includes the formulae used to calculate each goodness-of-fit measure. For example, for the model of the first choice of large-surface store, we have the following output (in which we show McFadden's Pseudo R-squared in bold face):

```
Goodness-of-Fit Measures
=========================================================================
Measure                  Value    Formula
-------------------------------------------------------------------------
Likelihood Ratio (R)     5541.3   2 * (LogL - LogL0)
Upper Bound of R (U)     19809
Aldrich-Nelson            0.5827  R / (R+N)
-------------------------------------------------------------------------
Cragg-Uhler 1             0.7525  1 - exp(-R/N)
Cragg-Uhler 2             0.7577  (1-exp(-R/N))/(1-exp(-U/N))
Estrella                  0.8057  1 - (1-R/U)^(U/N)
-------------------------------------------------------------------------
Adjusted Estrella         0.8024  1 - ((LogL-K)/LogL0)^(-
                                  2/N*LogL0)
McFadden's LRI            0.2797  R / U
Veall-Zimmermann          0.6995  (R * (U+N)) / (U * (R+N))
-------------------------------------------------------------------------
N = # of observations, K = # of regressors
-------------------------------------------------------------------------
```

As one can see, these are all measures of the difference between L, the likelihood of the fitted model, and L0, the likelihood of the null model.

We initially focused on McFadden's Likelihood Ratio Index (LRI, in the goodness of fit results above) or pseudo-R2, since it is a classic goodness-of-fit indicator that is designed to resemble the R2 index used in evaluating linear regression models. It produces a value between 0, the case where the likelihood of the fitted model is identical to the likelihood of the null model, and approaches one as the ratio of likelihood of the fitted model to the null model approaches infinity. We note that when compared with R2 statistics in linear models, the value of 0.2797 in the goodness-of-fit calculations above seems quite low. However, in chapter 5 of Domencich and McFadden (1975), the authors use results from simulation to show that seemingly low values of the McFadden's pseudo-R2 may actually correspond to high values of an R2 index calculated from a sum of squared residuals. In their simulations, a McFadden index of 0.30 corresponds roughly to an R2 calculated from squared residuals of over 0.60. We soon stopped using this index, however, for when we looked at the values of the McFadden LRI for various models of store choices, we found that this value varied according to our definition of the choice set for each individual, and did not represent the accuracy of the model's predictions well (Markley, 2006a).

Goodness-of-fit indicators would be meaningful if we were to use our discrete choice model in order to explain the behavior of individuals within our sample, the way it is often used in econometric papers. However, that is not our purpose. We wish to evaluate how well our model produces predictions of households whose behavior is unknown, and not simply how well it fits the data set used to calculate the parameters of the model. The problem with these indicators is that if we produce models that maximize goodness-of-fit without taking into account other considerations, we risk overfitting our model to our data.

For this reason, we do not focus on goodness-of-fit indicators, instead preferring to focus on the estimates of the parameters of our model. Here, we present the maximum likelihood estimates of our model of the first choice of large-surface stores, the values that are recorded in the file defined by the statement "ods output Parameterestimates = ParmEsts1".

We attach a great importance to the last column of Table 6 since this is essentially what we use in order to decide which explanatory variables to use in our model. This column represents the p-values resulting from the tests of the hypothesis that the corresponding true model parameters are zero. If the true model parameter in question were zero, then the parameter estimate divided by the standard error of this estimate (the t-statistic in the fifth column of Table 6) would follow an approximate normal distribution, provided the sample on which this estimate is based is sufficiently large.

The values in the third column of Table 6 form the vector $\hat{\beta}$ of estimated model parameters, which will be the basis for our model's forecasts on new data sets.

Table 6. Model estimations of the choice of the large-surface store that is the most visited amongst households in the Indre-et-Loire departments (small hypermarkets, HM, are the reference choice).

		Standard		Approx
Parameter	Estimate	Error	t Value	Pr > |t|
SM	-2.6523	0.1550	-17.11	<.0001
SMRankGE2	-0.7010	0.0793	-8.84	<.0001
SMRankGE3	-0.7951	0.1290	-6.16	<.0001
HMRankGE2	-0.2637	0.0748	-3.52	0.0004
HD	-4.8123	0.3000	-16.04	<.0001
HDRankGE2	-0.5205	0.1928	-2.70	0.0069
HDRankGE3	-0.3838	0.2585	-1.48	0.1376
XM	-11.1647	2.4160	-4.62	<.0001
XMRankGE2	-1.4753	0.1060	-13.92	<.0001
outside	-2.8539	0.1578	-18.08	<.0001
nostore	-5.1565	0.1974	-26.12	<.0001
disSM	-0.2300	0.0132	-17.37	<.0001
disHD	-0.1067	0.0163	-6.55	<.0001
disHM	-0.1647	0.008258	-19.95	<.0001
disXM	-0.0933	0.006741	-13.84	<.0001
surfSM	1.2718	0.0618	20.59	<.0001
surfHD	2.1797	0.2603	8.37	<.0001
surfHM	-0.1510	0.0210	-7.19	<.0001
surfXM	0.9808	0.2269	4.32	<.0001
GSVC99_1	0.3529	0.0757	4.66	<.0001
GSpol99_1	-0.8700	0.1119	-7.77	<.0001
gspol99_23	-1.1580	0.1176	-9.84	<.0001
FavCom	0.4692	0.1604	2.93	0.0034
DensPopu	-0.4372	0.0334	-13.09	<.0001

Estimates based on conditional logit using the MDC procedure.

We can therefore use this statistic to calculate the p-value (the last column of Table 6), or the estimated probability that a maximum likelihood estimate run on a data set resulting from the same data generating process that produced our sample will be further from zero than the one that we observed if the true model parameter in question were zero. For example, the estimated coefficient of the variable HDRankGE3 is -0.3838. If we generate a data set using the same data generating process that created our data set, then a p-value of 0.1376 indicates that provided the true value of the coefficient is zero (meaning that households do not differentiate the attractions of hypermarkets by their retail spaces) then there is a probability of 0.1376 that the estimate of the coefficient of HDRankGE3 that we calculate on our data set will be greater than 0.3838 or less than -0.3838.

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. If the p-value of a parameter is high, we consider it non-significant, since there is a high probability that we can generate a parameter estimate of the same size even when the true model parameter is zero. If it is low, then we consider it significant, and we are more certain that the parameter estimate represents a true effect. This is useful information, for in evaluating the parameters of our model, we can ignore those parameters that are non-significant, while trying to judge whether the parameters that are significant correspond to our beliefs about the relationship between the effects in question and the behavior we are trying to model.

However, our ultimate goal is not so much identifying the causes of consumer behavior (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 t-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 behavior of individuals in another data set.

The estimates of the parameters calculated using MDC 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 behavior 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 behavior we wish to predict, since their behavior is unknown. For this reason, we 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 behavior 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 judgment. In our case, we have confidence, 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) that showed 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 set follow the same patterns of behavior, 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 the true model parameter were zero, that the estimated parameter would be further from zero than the observed value. However, if we re-centered 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 if the estimated coefficient of HDRankGE3 in our model is 0.1376, and this 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 0.0688 that the coefficient of HDRankGE3 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 behavior 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 non-significant 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 variables 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 behavior 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 lends itself best to a logical interpretation. Table 6 represents the parameters estimated for the model of the choice of the large-surface store most visited after we have eliminated the variables that produced high p-values, so that all variables left have p-values less than 0.15.

We must now check to see that we can produce a satisfactory interpretation of our model coefficients. We see that the coefficients of the SM, HD, XM, and nostore are all negative. These coefficients represent the differences between the utilities of each choice of type of large-surface store and the utilities of small hypermarkets, all else being equal (note that small hypermarkets are the reference choice, since HM is not included as an explanatory variable in the model). The values of the coefficient of outside and nostore are not meaningful, since there are no retail spaces or distances associated with these alternatives, and so it is pointless to discuss the value of a utility "when all else is equal". The fact that the coefficient of SM is negative indicates that a household would prefer a small hypermarket to a supermarket if both stores were at the same distance from the household's home, were in the same type of commune, were of the same rank of distance from the household's home amongst stores of the same category, and had the same retail space. This interpretation is irrelevant, however, since hypermarkets and supermarkets cannot have the same retail space.

The interpretations of the other coefficients in this output are far more straightforward.

- We see that households prefer large-surface stores that have a lower rank of distance, all else being equal, as we can see by the fact that the coefficients of SMRankGE2, SMRankGE3, HMRankGE2, HDRankGE2, and XMRankGE2, are all negative. This means that households attach less utility to stores that are not the closest, than those that are.
- We see from the negative coefficients of the variables representing distance, and the positive coefficients of the variables representing retail space, that for all store types, aside from small hypermarkets, the greater the distance of the store from the household, the lower the utility, and the larger the surface area, the greater the utility. This is what we expect. What we don't expect is to see a negative coefficient of the variable surfHM which implies that for hypermarkets with less than 8000 square meters of retail space, the larger the store is, the less attractive it is. This does not agree with our original idea of how the people behave, and indeed, running our model on other departments and on the entire region yields positive values for this coefficient. This leads us to believe that within the Indre-et-Loire region, the larger hypermarkets within the class of small hypermarkets happen to be less attractive than the smaller hypermarkets, but this does not correspond to an overall trend in the comportment of French shoppers within the region.
- The coefficients of the variables GSpol99_1, and GSpol99_23 represent the utilities of selecting a large-surface store in an urban pole, or a large surface store in a commune classed as either monopolarized or multipolarized, compared with the utility of selecting a large-surface store in a nonpolarizeed commune, all else being equal. These coefficients are both negative, meaning that households, when faced with a choice of two stores of the same characteristics and the same distance from their homes, but one being in a polarized (and therefore more urban) commune, and the other within a nonpolarized (and therefore more isolated) commune, apparently prefer to visit the store in the nonpolarized commune. This could either be because stores in cities are harder to access due to the difficulties of going through traffic (although this effect should have at least somewhat been taken into account when we included a variable reflecting population density), or because households prefer the rural setting in which to do their shopping.
- We see that the coefficient of GSVC99_1 is positive, meaning that there is a slight preference for visiting stores that are in the inner city of an urban area, compared with stores in the suburbs. An advantage of locating a store in a city centre may be that households will more likely pass in front of the store as they go to work, or pursue other activities in the city.
- Population density is a dissuasive factor in store choice, which we assume is due to the greater difficulty in accessing the large-surface store due to associated traffic congestion.
- We have found that the fact that a large-surface store is in the same commune or department as the household in question, or is in a commune that has been found to be the most frequented by fellow residents of the household's commune, then it has a much higher utility.

In each of these cases, it would seem that we can interpret the coefficients of our variables in terms of the cost of accessing the large-surface store. Due to the smaller number of stores chosen for the second and third choices of large-surface stores, the coefficients of the explanatory variables for the model of these choices are more often non-significant (see Tables 7 and 8). However, we see that the same behavior patterns are present as in the first choice of large-surface store, which could allow us to tolerate the inclusion of some coefficients that have very low significance, by assuming that the behavior of individuals in the second and third choice of large-surface store is analogous to the behavior of individuals for the first.

E. Model Predictions

We have three models that produce three files of output, assigning a probability for each household of selecting each alternative presented to it. With a choice set of 12, this gives us 36 different values predicted per individual in our data set. With so many variables, it is difficult to summarize a household's behavior patterns, or compare it with that of other households, and so we need techniques to condense this information.

Table 7. Model estimations of the choice of the large-surface store that is the second-most visited amongst households in the Indre-et-Loire departments.

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
SM	-1.4364	0.1322	-10.86	<.0001
SMRankGE2	-0.4299	0.0936	-4.59	<.0001
SMRankGE3	-0.6220	0.1376	-4.52	<.0001
HD	-1.7175	0.2069	-8.30	<.0001
HDRankGE2	-0.5829	0.1462	-3.99	<.0001
XM	-10.7046	2.2141	-4.83	<.0001
XMRankGE2	-0.7349	0.0865	-8.49	<.0001
outside	-0.3433	0.0487	-7.05	<.0001
disSM	-0.1288	0.0133	-9.66	<.0001
disHD	-0.0923	0.0130	-7.09	<.0001
disHM	-0.0681	0.006023	-11.30	<.0001
disXM	-0.0515	0.004040	-12.75	<.0001
surfSM	0.7930	0.0721	11.00	<.0001
surfHD	0.3956	0.2303	1.72	0.0858
surfHM	-0.0819	0.0123	-6.69	<.0001
surfXM	1.0272	0.2057	4.99	<.0001
GSVC99_1	0.3885	0.0668	5.81	<.0001
gspol99_23	-0.2181	0.1147	-1.90	0.0574
FavCom	0.5665	0.1619	3.50	0.0005
DensPopu	-0.3382	0.0270	-12.54	<.0001

Estimates based on conditional logit using the MDC procedure.

Table 8. Model estimations of the choice of the large-surface store that is the third-most visited amongst households in the Indre-et-Loire departments.

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
SM	-0.5925	0.2214	-2.68	0.0074
SMRankGE2	-1.0529	0.1519	-6.93	<.0001
HD	-0.9620	0.1746	-5.51	<.0001
HDRankGE3	-1.1886	0.3367	-3.53	0.0004
XM	-14.6241	3.2284	-4.53	<.0001
XMRankGE2	-0.3151	0.1246	-2.53	0.0114
outside	0.4495	0.1559	2.88	0.0039
nostore	2.5449	0.1474	17.26	<.0001
disSM	-0.0714	0.0208	-3.44	0.0006
disHD	-0.0523	0.0165	-3.17	0.0015
disHM	-0.0343	0.0102	-3.38	0.0007
disXM	-0.0276	0.004903	-5.64	<.0001
surfSM	0.5024	0.1194	4.21	<.0001
surfXM	1.4146	0.3010	4.70	<.0001
gspol99_123	-0.2619	0.1155	-2.27	0.0234
DensPopu	-0.1199	0.0342	-3.51	0.0004
SameCit	0.3905	0.2570	1.52	0.1287

Estimates based on conditional logit using the MDC procedure.

In order to get an idea of the geographic layout of our model predictions, for every individual in our population, we create vectors containing the predicted probabilities of selecting every alternative for every choice of large-surface store. We then use data clustering in order to regroup survey sectors in clusters that have the lowest between-group variance possible. This allows us to identify areas in our sample where households have distinct behavior patterns.

We begin by taking the survey sectors in the population as our initial set of clusters, and use the Cluster and Tree Procedures in SAS in order to produce a dendrogram. For the Cluster Procedure, we require the averages and root-mean-squared areas of the vectors of probabilities in each survey sector. This can be calculated much more conveniently using the FastClus Procedure in SAS, than the Means or Summary Procedures, since it produces an output that is already adapted to the Cluster Procedure. The dendrogram is displayed in Figure 7.

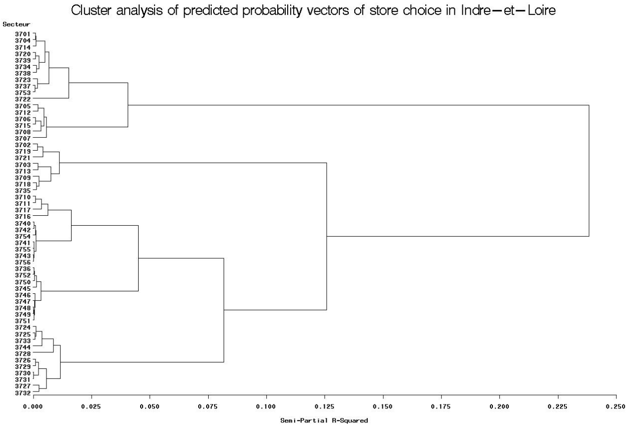

Figure 7 : Dendrogram of the 56 survey sectors in the Indre-et-Loire department (department 37) classified by the average of the probabilities of selection of each alternative for the three choices of large-surface stores.

At the beginning, we have 56 sectors which represent an R-squared of 0.690. When we arrive at the level where grouping clusters decreases the R-squared by more than 0.05, we have four clusters, and we see that by combining the two clusters that have the lowest between-cluster variance, we will reduce the total R-squared of the population by 0.082. The R-squared value of a division of the population into four clusters is 0.446.

The following four graphs (Figures 8 to 11) show the average predicted probabilities of each alternative and each choice of large-surface store for households in each of the four clusters generated using the cluster analysis. Our data clustering ought to ensure that these graphs are as distinctive as possible.

These four clusters can quickly be summarized. Cluster 1 represents households who don't have a dominant preference. Cluster 2 represents supermarket choosers: households who are much more likely to choose the closest supermarkets to their homes. Cluster 3 represents supermarket/large hypermarket choosers: households who are most likely to choose either large hypermarkets, supermarkets, or both. Cluster 4 represents small hypermarket choosers: households who tend to choose small hypermarkets. In Figure 12 we have mapped out the survey sectors according to the clusters in which they belonged. We also charted the distribution of supermarkets, hypermarkets and hard discount stores in the department.

We can see some geographic justifications for the different predicted behavior patterns. We see that the sectors in rural areas further from Tours and usually containing no more than one supermarket are often supermarket choosers in Cluster 2. With one exception, when these sectors far from Tours contain a hypermarket, or border a sector containing a hypermarket while not containing a supermarket itself, they tend to become small hypermarket choosers in Cluster 4. There is a cluster of sectors in the outskirts of Tours that contain a supermarket, but are within access of the large hypermarkets in Tours and are thus classified in Cluster 3, or households following a mixed supermarket/large hypermarket buying pattern. The last set of sectors represent sectors whose inhabitants do not favour any one store format, probably a result of the density of store choice, and the greater complexity of transportation patterns which increases the variability of store access times from household to household. Aside from these generalizations, there are some anomalies: notably a sector in the West of the department containing a small hypermarket yet classed as a sector of supermarket choosers, sectors just North of Tours classed as supermarket choosers despite being very close to the hypermarkets of Tours, and several sectors far from Tours whose predicted behavior patterns do not favor any one particular format.

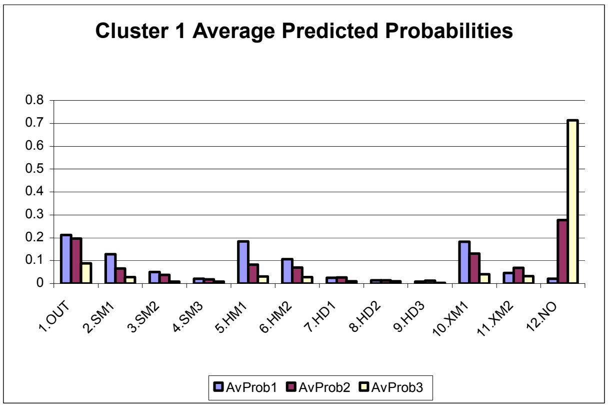

Figure 8 : Average predicted probabilities of households in survey sectors grouped in Cluster 1.

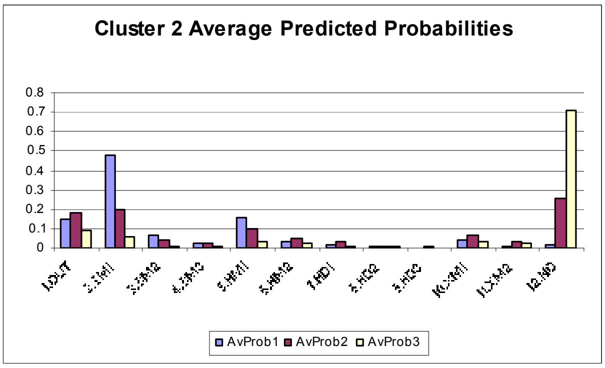

Figure 9 : Average predicted probabilities of households in survey sectors grouped in Cluster 2

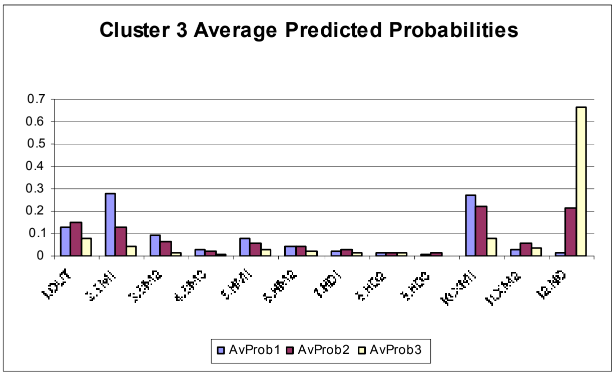

Figure 10 : Average predicted probabilities of households in survey sectors grouped in Cluster 3

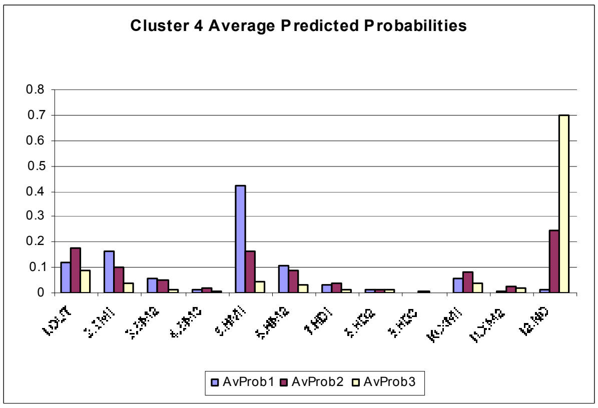

Figure 11 : Average predicted probabilities of households in survey sectors grouped in Cluster 4

The existence of hard discount stores in a survey sector does not seem to have a great influence in determining the cluster in which it is classified; this is a result perhaps of the choice of hard discount stores being less sensitive to distance than the choices of other types of stores. We see here that the probabilities of selection of large-surface stores in our simple model are largely influenced by the distribution of stores in the survey area, and follow common sense.

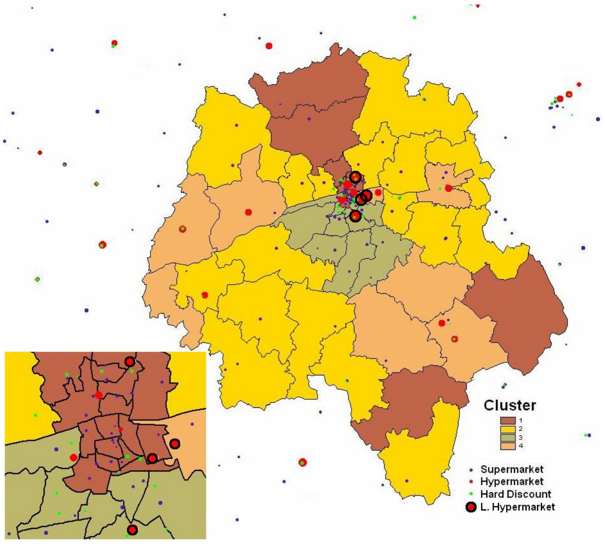

Figure 12 : Survey sectors in the Indre-et-Loire department, by cluster of predicted probabilities. The size of the points representing large-surface stores is proportional to their retail spaces.

IV. Validation

A. Test statistics.

Once we have developed a model of shopping behavior, we now evaluate it. One way to check the accuracy of our model is to compare the predicted probabilities with the actual choice of alternatives for the households in the sample. We can do this by calculating the average predicted probability assigned to the alternative that was actually selected by each household for each choice. Doing this over our entire set of 3968 individuals, we find that on average, the chosen alternative is assigned a predicted probability of 0.30.

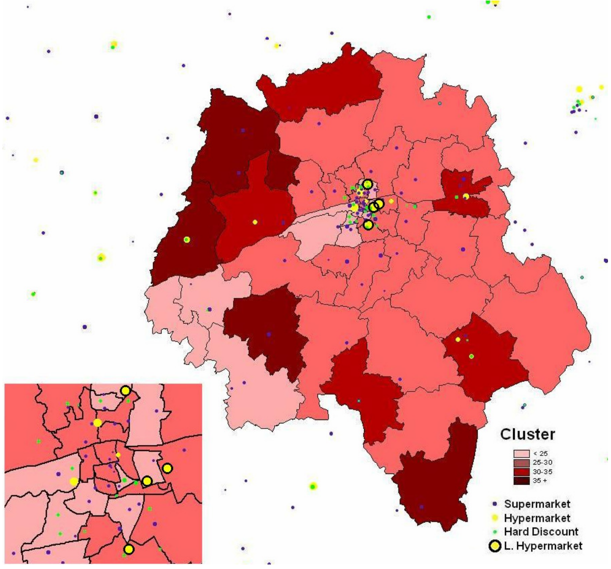

Figure 13 : Survey sectors in the Indre-et-Loire department by average predicted probability of selecting observed choice of large-surface store.

We see here (Figure 13) that for most survey sectors, the average predicted probabilities of the observed alternatives are about the same. Some of these lend themselves to easy explanations. We see that sectors with very high probabilities assigned to the selected alternatives are those sectors that are both far from Tours, and that have either only one store, or have a hypermarket that would presumably dominate the competition in the sector. Survey sectors with lower predicted probabilities of selecting the chosen alternatives are those that are on the outskirts of Tours, perhaps representing residents who are enticed away from choosing the store within their neighborhood by stores in the downtown area that could be closer to their workplaces.

What interests BVA in our modeling project is not so much predicting the behavior of an individual, but in predicting the aggregate behavior of populations. For example, we imagine two households in one survey sector: Household 1, who shops in Store A, and Household 2, who shops in store B. We could consider a model that predicted that Household 1 shopped in Store B, and Household 2 shopped in Store A as inaccurate, since neither prediction matches the actual choice of store for the household. However, a model that predicted that both households would choose Store A makes more correct predictions, yet we do not believe that it is preferable, since it did not accurately predict that one of the two household would choose Store A, and the other would choose Store B, as our first model did. We believe that from a retailer's perspective, obtaining the right number of clients visiting each store is more important than getting as many predicted stores matched with observed stores as possible. For this reason, we have developed an index of quality that would reflect this preference.

We begin by using the probabilities of selection predicted by our model to assign a store choice to each of our households. We considered assigning to each household the alternative corresponding to the highest predicted probability. However, this leads to biased results, as alternatives with high probabilities are assigned far more frequently than they are selected in the actual population. For example, all households with a predicted probability of 0.51 of selecting the closest supermarket will be assigned the closest supermarket, whereas according to our model, we ought to expect almost half of these individuals to choose another store type. The choices we assign are therefore the alternatives drawn at random with probabilities of selection corresponding to the probabilities predicted by the model. Since we are assigning three store choices to each household, we ensure that when a household chooses no store for one of its alternatives, it chooses no store on subsequent alternatives, and that no store choice is chosen more than once.

Once we have assigned an alternative to each household, we calculate the percentage of assigned choices of stores that can be matched with a household in the same geographic area observed to choose the same store. We shall call this index DA (for Drawn Alternatives). For example, suppose we have a survey sector in which 30 households choose Store A, 50 choose Store B, and 20 choose Store C, and we assign the choice of Store A to 40 of these households, Store B to 30, and Store C to 30. Then 30 of the assigned choices are choices of Store A that can be matched with an observed choice of Store A, 30 of our assigned choices are choices of Store B that can be matched with an observed choice of Store B, and 20 of our assigned choices are choices of Store C that can be matched with an observed choice of Store C. This gives a total of 80 households that can be matched one-to-one to a unique household in the population so that the assigned choice of one matches the observed choice of the other. We see that the number of households assigned a given store that can be matched with the number of households observed choosing the same store is simply the minimum of the two totals. The formula for calculating this value for geographic zone s is given by the following expression, where Aij is one if household i is assigned to store j and zero otherwise, and Oij is one if household i is observed to choose store j and zero otherwise:

$$DA(s) = \frac{\sum_{k} \min\left(\sum_{i\in s} A_{ik}, \sum_{i\in s} O_{ik}\right)}{\sum_{k} \sum_{i\in s} O_{ik}}$$

Once we calculate this value for each survey sector, we can take the weighted average of each survey sector to obtain a global index of model quality. The higher the number, the more the assignment of store choices will match the observed store choices. Taking the weighted average of these values of DA for each survey sector for the Indre-et-Loire Department, we arrive at a value of 0.769. This means that on average, for any given store, either the predicted number of clients coming from one survey sector will be roughly 77 percent of the observed number, or the observed number will be 77 percent of the predicted number. We calculate this figure by sector and presented the results in the following map (Figure 14). It is striking that the sectors scoring a higher value of the DA are not necessarily those that have high average predicted probabilities of selecting the observed choice of store, nor does this seem to correspond to the clusters of shopping behavior.

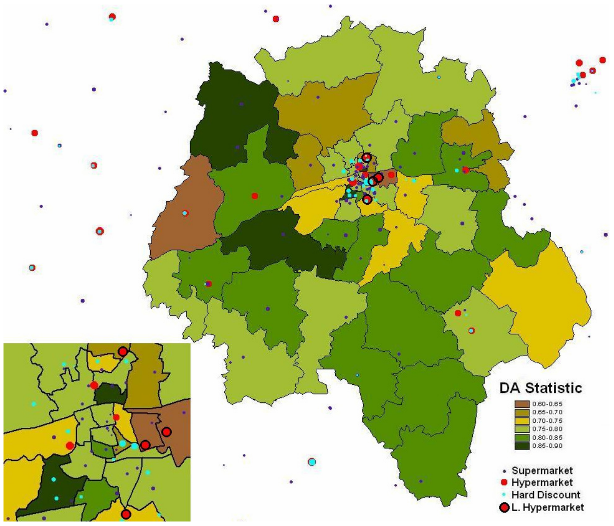

Figure 14 : The DA statistic for the Indre-et-Loire department

B. Testing on the Indre Department

One unfortunate problem here is that the way we validate our model is by comparing predicted probabilities to observed choices in a population where the behavior has been observed. However, observed behavior in this population was used in order to determine the predicted probabilities, and our model validation becomes circular. One way we can get around this problem is by using one set of known households for whom the choice is known in order to develop our model, and then we calculate the predicted probabilities of selecting each alternative on a different data set. Comparing predicted with observed values on this data set ought to pose no problem since the observed choices on the new data set are not used to generate the model's parameters.

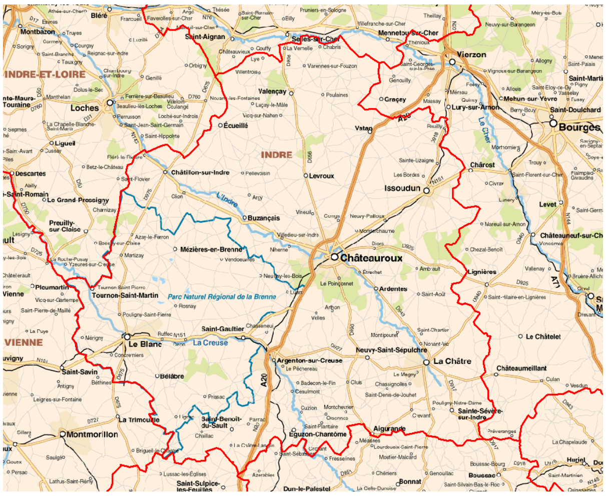

Figure 15 : Map of the Indre department of France (Department 36)

We have thus applied our model to the Indre department (36) neighboring the department Indre-et-Loire to the Southeast. This department is similar to the Indre department, having a mix of urban and rural settlement, and centered on one larger city. The difference here, however, is that its main city Châteauroux is far smaller than Tours. In general this department is much more rural and its inhabitants more isolated than those in Indre-et-Loire are (Figure 15).

We find that our DA for the Indre department, using the parameters of the model adapted to the Indre-et-Loire department is 0.747. This means that in going from a training set to a test set, we have not lost much predictive accuracy. The following map (Figure 16) shows the survey sectors in our survey by the values of the DA.

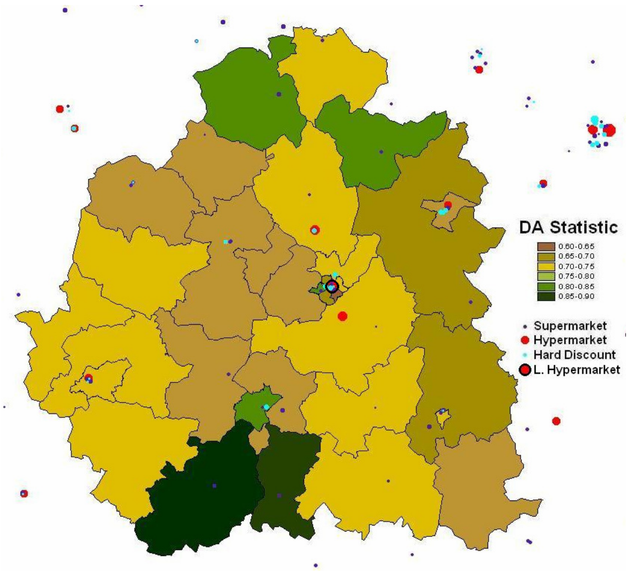

Figure 16 : Map of DA by survey sector in the Indre department.

V. Conclusions

We have developed a model of spatially-oriented discrete choice that was simple both conceptually and in terms of implementation, and that could be calculated rapidly. We have found that spatial data are indispensable to the modeling of store choice, first of all, since such data are necessary in order to determine a workable choice set to be defined for each household in our survey, and secondly because we have found that our model relies mostly on effects related to the geographic distribution of stores and households in our survey. In order to take into account these effects, we need to assign spatial co-ordinates to households and to large-surface stores and use these in order to create variables representing this distribution's effect on store choice. We can conclude that the effectiveness of our model depends upon the degree to which we can make our assigned geographic co-ordinates of households and stores precise.

After having reduced the choice set of our model to around 12 alternatives per household which contained the large-surface stores that were most likely to be chosen by the household, we constructed a model based almost entirely on spatial and geographic effects. The use of socio-demographic variables to represent taste variations was not used, since it had been determined that this did not lead to sufficient improvement in our model. The intrinsic appeal of stores could be accounted for by store type (supermarket, hypermarket, etc) and to a lesser extent its retail space. The utility of each store was determined first of all by the rank of its distance from the household, relative to other stores of the same type, then by its distance from the household's home, by the number of competitors it had in the same community, and finally, by the geographic characteristics of the community in which it was found. These effects were allowed to differ depending on the type of store in question. The model's parameters were restricted to main effects and were easy to interpret and judge as representing behavior patterns consistent with our expectations. Our choice of variables could be validated as the basis of a model used for prediction through the use of tests of significance, while global tests of goodness of fit were not used since they were not adapted to our purposes. By using a cluster procedure on the probabilities of selection predicted by our model, we were able to distinguish between the different general patterns of large-surface store choice, and found that these categories of predictions corresponded to the geographic features of our survey area in a logical manner. Our model was further validated through measures that contrasted the model's predicted probabilities of selection of alternatives with the observed alternatives selected. A measure of the accuracy of the predicted probability at an individual level could be illuminating, but we were more interested in comparing the number of households from the same survey sector choosing a given store, and the expected number predicted by our model. Although our model could not be used to predict the choice of a single individual accurately, the calculation of the "DA" led us to conclude that it could be reliable for the prediction of store clienteles, something that interested BVA. When we tested the stability of our model, by applying the same model parameters calculated on one department to another, we found that the model was very robust, thus allowing us to conclude that the original purpose of the project proposed by BVA, that is, predicting the behavior of households everywhere in France, based on data from one region, was substantiated.

We have thus created a simple and robust model of supermarket choice. There remain opportunities for improvement in this model, for a reader wishing to experiment with the data. We have already mentioned that a greater precision in our calculation of distances between households and stores can improve our model. Our geographic co-ordinates are already relatively precise, but we may be able to improve our model by replacing Euclidean distances between households and stores with estimated travel times. The model presented in this paper does not contain a lot of the variables that we have used elsewhere representing significant effects on supermarket choice. By adding detail to our model, one can gain small improvements in the quality of the model's predictions, but we do not believe that this will change the model's predictions to the point of radically altering the results we produced in this article. Another opportunity for improvement lies in adding complexity to the model structure. We assumed that error terms remained strictly independent identically distributed, rendering this model the simplest that could be specified. A more complete model may take into account spatial effects through the introduction of a spatial correlation of error terms (Dugundji and Walker, 2005). However, we have developed our model so as to minimize the necessity of recourse to such methods, by incorporating as many spatial effects as possible in the explanatory variables of the model, and in defining choice sets for each individual on spatial criteria. One possible way of improving our model that remains would be through the use of a Mixed Logit model in order to introduce a random variation of model parameters (Train, 2003).

Acknowledgements

I would like to thank my thesis supervisors Anne Ruiz-Gazen of GREMAQ and Michel Simioni of INRA as well as Dominique Haughton and the reviewers of CSBIGS for their helpful guidance, comments, and revisions. I also wish to thank my supervisors at BVA, Jean-Philippe Lesne and Hervé Tranger, who are responsible for the research project from which this article and the data used within it originated.

Correspondence: sebastien.markley@bva.fr

REFERENCES

Ben-Akiva, M., and Steven R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*, Cambridge, MA; M.I.T. Press.

Domencich, T. and Daniel L. McFadden. 1975. *Urban Travel Demand: A Behavioral Analysis,* North-Holland Publishing Co.

Dugundji, E.R., and Joan L. Walker. 2005. Discrete Choice with Social and Spatial Network Interdependencies, *Transportation Research Record*, 1921: 70-78.

Erdem, T., Susumu Imai, and Michael Keane. 2003. Model of Consumer Brand and Quantity Choice Dynamics under Price Uncertainty, *Quantitative Marketing and Economics* 1(1): 5-64

Hendel, I, and Aviv Nevo. 2006. Measuring the Implications of Sales and Consumer Inventory Behavior, *Econometrica*, 74(6): 1637-1673

Leboucher, Séverine. 2006. Qui sont les champions… du discount alimentaire, *Le Journal du Management*, February 15.

Markley, Sébastien. 2006a. Predicting Large-surface Store choice: A Trade-off Between Complexity and Computational Feasibility, Working Paper.

Markley, Sébastien. 2006b. Intégration de données spatiales dans la modélisation des choix discrets : applications aux modèles de comportements d'achats des ménages français. Doctoral thesis in progress, Université de Toulouse I.

Severin, V, Jordan J. Louviere and Adam Finn. 2001. The stability of retail shopping choices over time and across countries, *Journal of Retailing,* 77: 185-202

Smith, Howard. 2004. Supermarket Choice and Supermarket Competition in Market Equilibrium, *The Review of Economic Studies* 71(1): 235-263.

Train, Kenneth. 2003. Discrete *Choice Methods with Simulation,* Cambridge University Press, Cambridge UK.

Appendix: Survey Questions

(Rough partial translation from French)

Hello, my name is NAME OF SURVEYOR of the BVA Institute.

We are undertaking a study concerning consumer behavior. This study has been commissioned by the Chambers of Commerce of NAME OF MAIN CITY IN DEPARTMENT. This study aims to determine where you buy different types of products. We look to understand the way commerce functions in the department better in order to try and adapt its evolution to the needs and practices of consumers.

Your home has been selected in order to provide a response representing a household to a questionnaire by which we wish to know which stores you visit in order to buy food products and non-food products as well as some information about your household.

This survey will last 30 minutes. I wish to interview the person who normally does the purchases for the household. Will you accept to participate in this survey? Would you accept to respond to certain questions?

SURVEYOR : IF THE PERSON IS NOT AVAILABLE, MAKE AN APPOINTMENT

- □ Yes, I will respond now
- □ No, I'm not interested
- □ No, I don't have the time
- □ No, I do not to give information about my household.
- □ What is the purpose? Who is ordering the survey?
- □ What will this bring me?

SCREENING (Questions related to confidentiality)

Car1: I will begin by asking a few questions about characteristics.
Are you single, or in a couple?

□ Single □ In a couple

Car2: Do you have an occupation, or no occupation (unemployed for more than a year, retired)?

□ Occupation □ No Occupation

Car3: Does your partner have an occupation or no occupation ?

□ Occupation □ No Occupation

We will begin by discussing your purchases of food products, beginning with purchases in large-surface stores. During the survey, we will search for the stores that your visit in a list that includes all commercial establishments in the region. This search will be done with your help, and with some cartographic tools I have in front of me. The more precise you are in your explanations, the more efficient we can be.

P1: Food purchases in Large-Surface Stores.

A1A. For your food purchases, what large-surface store (hypermarket, supermarket, or specialized large-surface store) do you visit the most often? The large-surface store the most often visited.

THE MOST OFTEN VISITED LARGE-SURFACE STORE IS NOTED USING GEOCATI

A1A. How often do you go to STORE INDICATED IN A1A ?

SURVEYOR: LIST

□ Several times a week □ Once a week
□ 2 or 3 times a month □ Once a month.
□ Less than once a month □ Don't know

A1B. Which products do you normally buy in STORE INDICATED IN A1A?

SURVEYOR: LIST (maximum 6 responses)

□ Breads and pastries □ Fresh fruits and vegetables
□ Meats and poultry □ Fish and seafood
□ Frozen foods □ Spices, creams, other food products, and maintenance products
□ None of the above

A1A. For your food purchases, in which large-surface food store (hypermarket, supermarket, or specialized large-surface stores) do you go the most frequently? The large-surface store the second-most often visited.

THE SECOND-MOST OFTEN VISITED LARGE-SURFACE STORE IS NOTED USING GEOCATI

A1A. How often do you go to STORE INDICATED IN A1A ?

SURVEYOR: LIST

□ Several times a week □ Once a week
□ 2 or 3 times a month □ Once a month
□ Less than once a month □ Don't know

A1B. Which products do you normally buy in STORE INDICATED IN A1A?

SURVEYOR: LIST (maximum 6 responses)

□ Breads and pastries □ Fresh fruits and vegetables
□ Meats and poultry □ Fish and seafood
□ Frozen foods □ Spices, creams, other food products, and maintenance products
□ None of the above