Modelling Shopping Frequencies

Tomas Svensson and Mattias Haraldsson
Swedish National Road and Transport Research Institute
SE-581 95 Linköping

Efficiency Implications of Households' Choice of Stores in Grocery Retailing

Introduction

The structural changes in retailing towards out-of-town establishments during the latest decades have profoundly affected the attractiveness of towns and cities and the urban transport system. Increasing car usage and dependency, decreasing production of retail services in residential areas, degeneration of inner cities and costly revitalisation programmes, reduced demand for public transport with negative consequences for its quality, are some of the factors that can be observed and, in part, linked to the changes in retailing. (Boverket, 1999) Increasing access and usage of the car, in combination with a historic record of accommodating town planning, has enhanced the exploitation of scale economies within distribution firms. (Svensson, 1998) As a reaction to this development, many countries in Europe have implemented policies and planning guidelines with a more restrictive attitude towards out-of-town establishments in retailing. This has been done, however, without an in-depth economic analysis capable of ranking different goods distribution systems (transportation, storing, handling etc.) in urban areas according to their overall socio-economic efficiency, when relevant externalities, distribution costs in the informal sector etc., are incorporated in the analysis.

There is a strong ambiguity concerning the overall efficiency of the distribution system when the costs associated with, for example, travelling and purchasing performed by households are treated as an integral part of the total distribution cost function. This paper will briefly discuss the research agenda in this field with examples and applications from the grocery sector, where wholesaling and retailing of groceries are defined as parts of an integrated urban production system. The focus of the research programme, which is partly based on surveys, is on modelling the consequences of different types of store locations in cities, and identifying the relevant external costs associated with a decreasing production of retail services in residential areas. (Svensson, 1999)
The main part of the paper is devoted to a discussion of some recent findings from the research performed by the authors into the efficiency implications of households` choice of stores in grocery retailing.

Research questions

The research questions in part dealt with in this paper concern the overall economic efficiency of grocery distribution systems in towns and cities. This main focus is in this paper linked to the possible identification of potential "social dilemmas" associated with structural changes in grocery retailing on a local level.
The aim of ongoing research is to bring together this perspective with others in an integrated and comprehensive analytical framework, as well as to identify relevant questions for further research.

Theoretical and methodological considerations

The research is based on economic theory, where distribution of goods is conceptualised as an integrated system with firms and households as the relevant actors performing distribution activities associated with different cost components. The nucleus of the system is the costs of transporting, storing and handling goods, activities carried out by production firms, distribution firms, i.e. wholesalers and retailers, and household members. Institutional and organisational barriers between firms and households, formal markets and informal sectors etc., must be handled in the analysis, where the system as a single entity defines the analytical framework. This approach has been developed in earlier research performed by the authors of this paper. (Svensson, 1998) The driving forces behind and consequences of out-of-town establishments in retailing are frequently analysed and debated, with the focus on car usage and environmental effects, consequences for other stores and inner-cities, price-levels and competition and so on. (See for example Forsberg et. al. 1994, Svensson, 1998, Hellberg, 2000, and Haraldsson & Svensson, 2000)

However, the strategy in current research, which in part is discussed in this paper, is to integrate several relevant aspects of the structural changes in grocery retailing into one integrated and comprehensive framework. This line of investigation is of course inspired by other attempts in the research field such as the CREATE model, i.e. combined retail, economic and transportation evaluation, discussed in DETR(1998). The general analysis starts with the most obvious and accessible cost items, such as in-firm costs registered by price levels for different categories of stores and generalised transportation costs for car traffic. (Haraldsson, 2000a) The data can be collected from governmental bodies, calculation recommendations and appraisals for public infrastructure investments, earlier research etc. When this possibility is exhausted, data has to be generated from primary sources such as questionnaires to households, which are used as data source in the study that is discussed below. The ultimate objective is to construct a model that is capable of handling all relevant costs in one single analytical device. (Haraldsson and Svensson, 2001) Although this target has not yet been reached, setting the agenda and pointing out the directions for future research can be a valuable support for planning authorities and policy makers. (Svensson, 1999)

Decreasing production of retail services as a social dilemma

According to Ostrom (2000), a social dilemma can be broadly described as a situation where benefits achieved from the consumption of a public good are hindered because the unique

equilibrium for rational egoists in a single-shot game is to contribute, or pay, nothing at all. This means that a Pareto-sanctioned transaction that would yield a social payoff way above every actor's own contribution is not accomplished and results in a social efficiency loss.
The question is if this line of argument has any bearing on the social and economic consequences of the structural changes towards larger and more distant food stores in grocery retailing. The answer should in principle be yes. (Forsberg, 1997) The problem can be discussed and conceptualised using a household's choice of stores when purchasing groceries. Assume that a household originally can use only one store, which is located in the residential area. Assume further that a new out-of-town supermarket is established, say 10 kilometres from the residence of the household. Definition, in accordance with the traditional economic cost minimising principles, of the factors that should decide which store to use and to what extent is possible but somewhat tricky. (Svensson, 2001 and Haraldsson, 2000b)

However, the important conclusion from that exercise is that a "rational egoist" household has no reason to let the risk of a vanishing "accessibility component", caused by a deteriorating production of retail services in the residential area, influence the decision. The reason for this is of course that one isolated urban household can not by itself influence the volume and quality of the local retail service. If this "accessibility component" were internalised with a positive value, i.e. with a guaranteed outcome associated with a certain behaviour, a rational household would buy more in the local store, or participate in some other institutional arrangement with the aim of keeping or improving local retail services.

Empirical tests of social dilemmas in grocery retailing

In order to find if there is any empirical equivalence to the theoretical discussion above, a pilot study was carried out. (The study is discussed at length in Svensson, 2001) This was done by means of a postal questionnaire with hypothetical questions about households` choice of stores, together with questions about socio-economic conditions and actual purchasing behaviour. The questionnaire was sent out to 400 individuals, living in the same district in the small Swedish town of Motala. The town has ca 30 000 inhabitants. Approximately 220 individuals answered the questions and returned the questionnaire. The hypothetical questions were constructed as choices of shares of the total amount of groceries purchased by the household in two different stores: a local neighbourhood store and a distant out-of-town store. The local store was more expensive than the out-of-town establishment. The price differences were calculated from the empirical cost functions estimated in Haraldsson (2000a). The respondent could choose to buy 0%, 10%, 25%, 50%, 75% or 100% of the groceries in the local shop, and the remainder, of course, in the larger store.

After a first round of questions options were altered. The consequences of different purchasing shares were described in a table, with the assumption that everybody in the area would choose the same alternative as the respondent. It is plausible to assume that the distant large store is unaffected by the behaviour of the residents in the particular area, but that the outcome concerning the local store is directly in proportion to the chosen share. Larger share means lower prices and higher quality in other respects, particularly as regards the range of products, in the local shop. The share "0%" implies, of course, that the local store is forced to close. The share "100%" implies that the local store is unaffected by the arrival of the new out-of-town establishment. When the respondent had been informed of the consequences, the same question was asked again: "How much groceries do you want to buy in the local store?"

This means that the respondent now chooses a hypothetical scenario for the level of retail service in the area, knowing that everyone else will behave in the same way. The accessibility component discussed above is now internalised in the household decision making process.

Thus, we obtain "before" and "after" scenarios, which can be compared. This is done in the following figure:

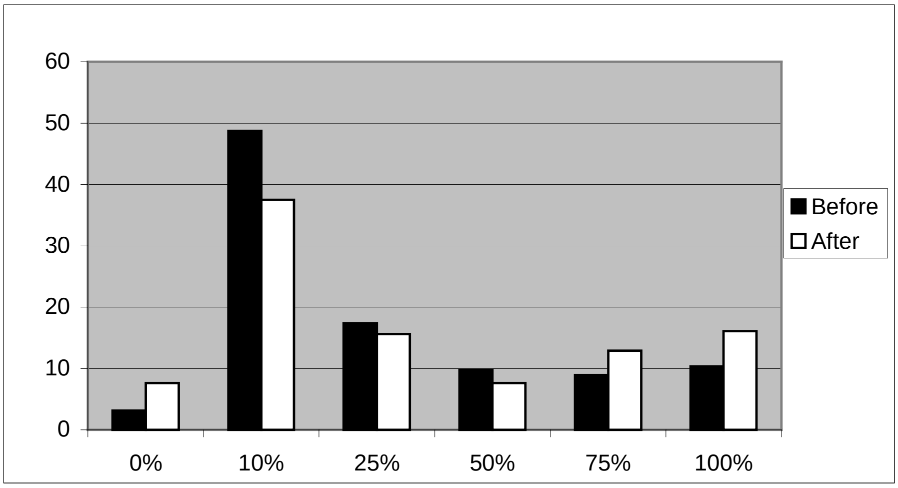


Figure 1. Distribution of shares in the local store, percentages

The columns show, in percentages, the shares chosen by the interviewees in the survey "before" and "after" the information of the consequences for the local store. For example: 49% of the respondents "before" and 37% "after" chose the share 10% in the local store. There is a clear overall tendency in favour of the local store in the "after" scenario. The average market share for the local store increases from 31% "before" to 39% "after". The increase is statistically significant ($p<0,05$). Around 37% per cent of the individuals choose to buy half or more of their groceries in the local store in the "after" case, compared with 29% "before". The result implies that the hypothesis about social dilemmas in this field is relevant, but has to be scrutinised further in forthcoming research.

A scattered preference structure

The distributions of shares, shown above, were analysed with the aim of finding the reasons behind the actual choice and the differences between "before" and "after". It was not possible to explain this difference with regression models containing significant coefficients. Although the difference between the two distributions is statistically significant, most of the interviewees preferred to choose the same alternative in the two cases. The research effort was therefore directed towards the distributions as such. To compensate for the rather small sample (400 questionnaires, with an answering frequency of ca 56%) "pooled" regression models were used, which means that the data described in the two distributions above were combined into one data set. (Greene, 1997 and Gujarati, 1995). A dummy variable with the value 1 for the "after" observations, and the value 0 for the "before" observations, was incorporated in the model.

In a first step only socio-economic variables and the dummy variable were used in the models. The determination coefficient was low in all specifications, less than 0,1, but the F-value was significant and some of the explanatory variables were estimated with significant coefficients: gender, income, age and numbers of household members. Lower income, lower ages, fewer household members, and if the interviewee is a woman, are factors that increase the share of groceries purchased in the local store.

The following table shows one of the estimated models:

Table 1. Pooled regression model: the share of groceries bought in the local store as a function of socio-economic variables and a before/after dummy

Variable	Constant	Gender (male)	Income	Household members	Before/after
Coeff.	0,556	-0,067	-0,033	-0,050	0,073
t-value	10,53	-2,15	-2,22	-3,10	2,35
R^2-adj.	0,07				
F-value	7,84				

By adding variables which describe the actual purchasing behaviour, such as frequencies, modes of transportation, actual store category used by the household etc., the determination coefficient in the regression models was improved dramatically, up to 0,4 in some specifications. These variables are of course not explanatory variables in a theoretical sense, only based on the plausible assumption that there is a correlation between revealed and stated preferences. However, problems with correlated explanatory variables and low variation resulted in some problems concerning the construction of the models when, for example, the significance of a particular coefficient was dependent on the model specification. The following table shows the combinatory model that resulted from an attempt with manual stepwise elimination of the explanatory variables:

Table 2. Pooled regression model: the share of groceries bought in the local store as a function of socio-economic variables, a before/after dummy and actual purchasing behaviour

Variable	Constant	Age	Before/ after	Frequency A-store
Coeff.	0,452	-0,024	0,084	0,010
t-value	6,83	-2,23	2,92	2,78
Variable	**Cycling A-store**	**Cycling C-store**	**Walking A-store**	**Walking C-store**
Coeff.	0,232	0,078	0,154	0,113
t-value	3,44	2,02	2,48	2,37
Variable	**A-store is an out-of-town store**	**B-store is an out-of-town store**		
Coeff.	-0,212	-0,162		
t-value	-5,59	-4,85		
R^2-adj.	0,31			
F-value	16,03			

The table needs some clarification. The respondents were asked about their actual buying behaviour in three different grocery stores, with free options regarding which store to choose. The purpose with that exercise was of course to capture differences between, for instance, main and complementary purchases, which are often carried out in different stores with different errand combinations etc. The three different categories, or levels, of stores is labelled A, B and C to simplify the description and analysis. The age coefficient is enlarged and describes the consequences for the share bought in the local store due to a ten-year increase in the respondents' age. "Frequency A-store" is measured on a monthly basis. The other variables in the model are dummy variables such as "before/after", see above, variables describing the mode of transportation most often used when travelling to the stores, and variables describing if the respondents` actual A- or B-store are out-of-town establishments. The logical variable gaps, such as "Cycling B-store" for instance, are omitted due to lack of statistical significance. The dummy variable "before/after" was significant in all specifications which, again, proves that the difference between the "before" and "after" case is statistically significant.

One complementary alternative to regression models when analysing the preference structure is to divide the material in half, those in favour of the out-of-town establishment in one, and those who prefer the local store in the other. This was done by letting those individuals who want to purchase 75% or more of their groceries in the distant store form "Group 1", and those who want to purchase at least half of their groceries in the local store constitute "Group 2". Group 1 is composed of ca 62% of all the individuals in the sample, and Group 2, consequently, of 38% of all the individuals. When the relative frequency distributions between the two groups were tested with the Wilcoxon Signed Ranks test (Aczel, 1996), the following socio-economic variables showed significant differences: gender, driving licence, car access or ownership, type of housing, and income. The different percentages of how men and women have chosen, so to speak, between the two groups are displayed in the following figure:

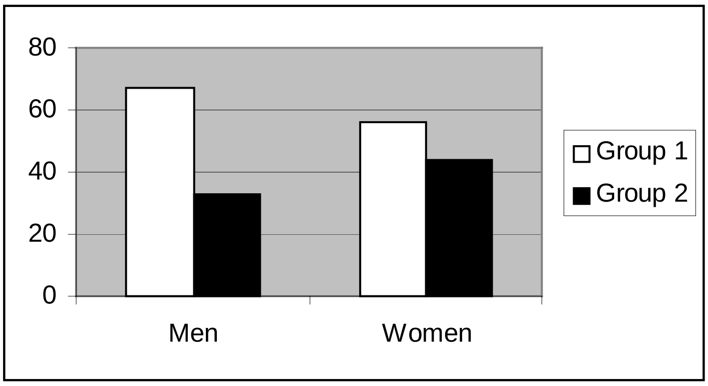


Figure 2. Men and women in the two groups, percentages

Group 1, i.e. those who prefer the distant store, is the majority's choice among both men and women. But the relative differences between men and women concerning the preferences for the two stores are evident. Two thirds of all men prefer the distant store. The corresponding proportion among women is only 56%. The relevant preferences about the two types of stores, and the following consequences for the purchasing behaviour in general, seem to be more evenly distributed among women. The following figure shows the distribution of different types of housing in the two comparison groups:

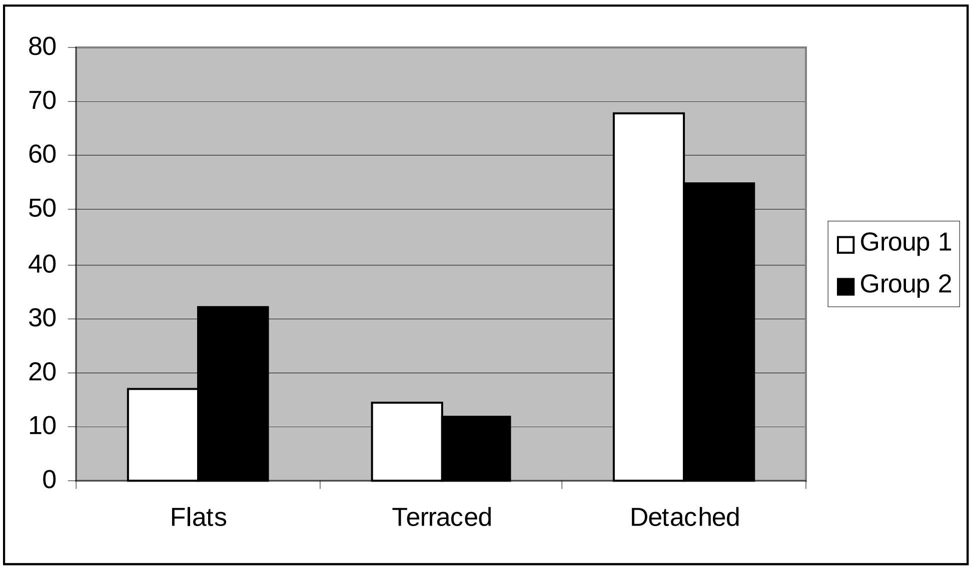


Figure 3. Type of housing in the two groups, percentages

The proportion of the respondents living in detached housing is clearly higher in the group that prefers the distant store, and the proportion living in flats is correspondingly lower. Furthermore, the proportion of driving licences and car access or ownership is significantly lower in the "local store group" than in the other group. The respondents in the group that prefer the local store have also a significantly lower income. This is somewhat unexpected considering the lower price levels in out-of-town stores, but corresponds with the results from *Table 1*.

Conclusions

The results are generated from a pilot study and should be interpreted with care. The result indicates, however, that individuals living in the same residential area have varying preferences about the demanded level of retailing service in the neighbourhood. Some prefer local stores and are, hypothetically, willing to pay for this service, and some prefer to drive their car to more distant stores and take advantage of lower prices. By using regression models, and other methods, it is possible to link these differences in preferences to socio-economic background variables. Gender, income, number of household members, and age can to some extent be used as explanatory variables in regression models. The probability that a younger woman in a single-person household without a car and driving licence, living in a flat, and with an income lower than average, will prefer the local store in the hypothetical questions is rather high. The statistical opposite to the woman, so the speak, is a middle-aged man in a household with many members, who lives in detached housing, owns a car and has a high income.

There is, however, a large variation in the data that can not be explained by the models. By adding variables describing the respondents` actual purchasing behaviour, the technical explanatory capabilities of the models are improved. This is, not surprisingly, caused by the obvious correspondence between stated and revealed preferences, even if the "revealed" preferences in this particular case are filtered through a postal questionnaire.

But it is clear that the actual structural situation in grocery retailing in the investigated area, and elsewhere, accommodates the preferences of those who prefer to do most of their grocery shopping in out-of-town establishments. The large stores attract customers from the entire city, which implies a diminishing market for the smaller local stores in residential areas, and a correspondingly decreasing production of local retail services. If individuals with more similar preferences lived in the same residential areas, the potential for a higher level of local retail service would have been larger.

This is of course not easy to accomplish, and might also lead to negative consequences for other aspects of society. But there is no doubt that more research efforts should be devoted to the efficiency losses that are caused by an uniform town planning, without differentiation and based on average preferences. (Jansson, 1996) The problems due to inefficient low production of local public goods are largely the same in other related areas such as the balance between individual car accessibility and environmental effects in urban settings, just to mention one of many relevant parallels (Svensson, 2000a and b, and Grudemo & Svensson, 2000)

To sum up, it seems that this pilot study has identified two kinds of efficiency problems which must be addressed in further research: social dilemmas impacting on the choice of which store to use, and consequences of geographically scattered preference structures among neighbours. There is, however, an obvious need for further research in this line of investigation.

References

Aczel, A(1996), *Complete Business Statistics*, 3rd ed. Irwin.

Boverket(1999), *Handeln i planeringen*, Karlskrona.

Department of Environment, Transport and the Regions, DTR(1998) *The Impact of Large Foodstores on Market Towns and District Centres*, London.

Forsberg, H et al(1994), *Effekter av externa köpcentran*, STACH Rapport 1994:1, Chalmers tekniska högskola, Göteborg.

Forsberg, H(1997), *Konsumenternas valmöjligheter vid dagligvaruinköp- Speglade genom rörlighet och tidsanvändning*, Rapport 1997:18, Konsumentverket, Stockholm.

Greene, W. H(1997), *Econometric Analysis*, 3rd ed. Prentice Hall International Editions.

Grudemo, S. och T. Svensson(2000), *Balans i avvägningen mellan biltillgänglighet och god miljö- slutrapport*, VTI rapport 455, Statens väg- och transportforskningsinstitut, Linköping.

Gujarati, D. N(1995), *Basic Econometrics*, 3rd ed. McGraw-Hill International Editions.

Haraldsson, M(2000)a, *Distributionsekonomisk optimering av dagligvaruhandelns lokalisering- förstudie till ett forskningsprojekt*, D-uppsats i nationalekonomi, Ekonomiska Institutionen, Linköpings universitet.

Haraldsson, M(2000)b, *Konsumentens val mellan olika butikstyper- en mikroekonomisk problemanalys,* PM, Statens väg- och transportforskningsinstitut, Linköping.

Haraldsson, M och T. Svensson(2000), *Förslag till analys av skillnader i livsmedelspriser mellan orter,* Bilaga 2 i Förslag till prisundersökningar och analys av prisskillnader- Redovisning av ett uppdrag från regeringen, april 2000, Konsumentverket, Rapport 2000:8.

Haraldsson, M. och T. Svensson(2001), *Economic analysis of out-of-town establishments in the retail industry,* pp.245-258 in Sucharov, L.J. and C.A. Brebbia (eds.) ”Seventh International Conference on Urban Transport and the Environment, Urban Transport VII”, WIT Press, Southampton and Boston.

Hellberg, C(2000), *De externa etableringarnas framtid- en samhällsekonomisk analys av behov och möjlighet till politisk styrning och planering*, D-uppsats i nationalekonomi, Ekonomiska Institutionen, Linköpings universitet.

Jansson, J. O(1996), *Transportekonomi och livsmiljö,* Stockholm.

Ostrom, E(2000), Collective Action and the Evolution of Social Norms*, Journal of Economic Perspectives*, vol.14(3), pp.137-158.

Svensson, T(1998), *Dagligvarudistributionens strukturomvandling- drivkrafter och konsekvenser för städers utformning och miljö,* diss. tema Teknik och social förändring, Linköpings universitet.

Svensson, T(1999), *Samhällsekonomisk analys av dagligvaruhandelns strukturomvandling,* Projektbeskrivning, Statens väg- och transportforskningsinstitut, Linköping.

Svensson, T(2000)a, *Biltillgänglighet och bilbegränsningar i innerstäder och bostadsområden,* VTI meddelande 893, Statens väg- och transportforskningsinstitut, Linköping.

Svensson, T(2000)b, Balancing car accessibility and good urban environment, pp. 99-110 in *Discovering Local Transport Plans and Road Traffic Reduction- Proceedings of Seminar C held at the European Transport Conference, Homerton College, Cambridge, 11-13 September 2000*. Volume P438. London 2000.

Svensson, T(2001), *Hushållens inköpsmönster och val av dagligvarubutiker- En förstudie,* VTI notat 6-2001, Statens väg- och transportforskningsinstitut

Modelling the Frequency of Shopping Trips using Poisson Regression Analysis

Introduction

The demand for transportation services is generally considered to be induced from an underlying activity, such as commuting to work, performing a leisure activity, visiting friends or relatives, or shopping. Following this line of argument, transportation from origins to destinations is complementary with respect to the activities that are linked together and should not be treated as isolated phenomena. The models traditionally used when analysing the demand for travelling, however, do not incorporate these real conditions in the analysis due to applied assumptions and preferred model design. But the growing research interest in this area and the recent development of activity based models provide the stimulus for an attempt to model the frequency of different kinds of activities such as shopping trips. (See the discussion in the proceedings from "Activity-Based Travel Forecasting Conference 1996").

Travelling for shopping purposes generates large flows of traffic in cities, a fact accentuated by the ongoing trend towards out-of-town establishments in retailing and the increasing use of the private car. Analysis of the frequency of shopping is important because of its consequences for the total traffic volume. Lowering the frequency will reduce the total amount of shopping trips and release scarce capacity in the urban transport system, which can be used for other purposes. Moreover, the results will be of interest when estimating the potential for shopping on the Internet. The consumer demand regarding the density and regularity of delivery schedules will have a vast influence on the economic efficiency of small-scale delivery services in residential areas.

The data set that will be used in this analysis is from a case study of the consequences of large out-of-town establishments in grocery retailing in the Swedish middle-sized city of Linköping. (Svensson, 1998) The registered purchasing frequencies are of course not unconditional. A shopping trip is carried out when the needs outweigh the generalised costs of travelling to the store, which in turn is a function of the physical structure of the distribution system, e.g. store location and integration with the transport system. The data set contains information about households' shopping behaviour for two different periods: before and after the new establishments. The total number of stores were approximately the same before and after, but the functions of the stores (main and/or complementary purchases etc.) changed as a result of the new large establishments, which altered the frequency of visiting particular stores.

It has been shown elsewhere that the explanatory variables used in the analysis explain the total amount of groceries purchased by a household at a satisfactory level. Svensson (1998) estimates an ordinary multiple regression model where the dependent variable is defined as the total volume, in monetary units, of groceries purchased by a household during the period of a month. The model shows, not surprisingly, that the volume purchased increases with the number of household members. Car access and income, both with estimated regression coefficients with positive signs, increase the volume purchased. Different types of housing such as flats, terraced and detached housing were also used as explanatory variables.

Households living in flats in multi-storey buildings purchase a smaller amount of groceries than households living in terraced and detached housing. The influence of variables such as car access for the particular household and type of housing, is explained by the fact that these characteristics raises or lower the generalised cost for the household in terms of transportation, handling and storing. But when the shopping and travelling frequencies are considered, it is harder to have any strong presumptions about the signs and magnitudes of the regression coefficients. Retired or unemployed individuals have possibly less restrictive time budgets, which will induce higher frequencies. Moreover, it seems plausible to assume that car holders purchase and transport larger volumes when shopping than individuals without a car, which tends to lower the frequencies. On the other hand, it is possible that the behaviour linked with shopping for complementary purposes will compensate for the consequences of other characteristics.

Aim and scope

The aim of this paper is to explain and quantify the differences in purchasing frequencies that can be observed between individuals and households. We will in particular discuss whether the Poisson regression approach can enrich the analysis and generate statistically significant coefficients that can be interpreted in a logical manner.

The data set and former research

The data used in the analysis was collected during the middle of the 1990s, and constituted some of the empirical material elaborated in the doctoral dissertation *Structural change of food distribution- driving forces and consequences for urban design and environment* (Svensson 1998). A case study with the character of a before-and-after-study was performed with the aim to analyse the consequences of large out-of-town establishments in grocery retailing in the Swedish middle-sized city of Linköping, a city with around 130 000 inhabitants. In 1993 and 1994 three large stores with a total selling area of approximately 17 200 square metres, were established in the Tornby industrial area on the outskirts of the city. A large part of the new stores was used for selling food and groceries. To obtain data about the changes in shopping patterns as a result of the Tornby establishment, a survey of shopping behaviour was carried out. This was done by means of a questionnaire which was sent to 2 000 households in Linköping. Approximately 72 per cent of the households returned the questionnaire with data that could be used in the analysis. Data was also collected from all the stores used by the households in the sample.

It was found that the three new large stores in Tornby quickly absorbed as much as 36 percent of the grocery retailing market in Linköping. In the short run the total consumption of food and groceries is rather inelastic, which means that the business of new and successful establishments is directly reflected in falling sales and losses for the original stores. The case study shows that all existing food stores in Linköping were influenced negatively by the new "hypermarkets" in Tornby, but the small stores used for complementary shopping were relatively less affected than the larger supermarkets in district centres. These larger stores had previously been used to a large extent for main purchases also by customers who use the private car when shopping.

The case study also shows that the shopping behaviour in an out-of-town establishment is quite different from the shopping behaviour in other food stores. The shopping trips are not as frequent, the volume purchased is larger, and about 90 percent of the households use cars when buying groceries in the out-of-town stores. The households seldom combine shopping with other purposes when travelling to the hypermarkets. In the regression analysis it could be seen that the share of the total shopping by individual households in the out-of-town establishments is positively related to household size and car availability, and negatively related to distance from home and income. The data collected was also used to calculate the total additional car mileage of households in Linköping caused by the new establishments. The result of the calculation was an increase of more than 4 million car kilometres per year, which corresponds to a relative increase of 50 percent in the total car mileage for everyday commodity shopping.

The questionnaire used for collecting data contained questions about shopping frequencies for three different grocery stores, stated by the households as the most frequently used stores and labelled A -, B - and C - stores. For our purpose the intervals used in the questionnaire must be transformed into quantitative numbers, which is a rather simple calculation that will not cause any biases in the data. The i:th observation is a pair of non-negative integers (p_i, q_i), where $p_i \leq q_i$. The total purchasing frequency is obtained by summation of the limits of the intervals for all the three categories of stores, as described in (1):

$$(1) \quad (p_i, q_i) = \sum_{j=1}^{3} (p_{ij}, q_{ij})$$

The relative frequencies that are calculated in Svensson (1998) can with some approximations be recalculated in order to define the average frequency. By substituting unique values for the intervals it can be shown that the average consumer purchases groceries 16 times per month. (Haraldsson, 2000) The research question tackled in this paper is to explain the differences between individuals and households, by considering the frequency used when they purchase groceries.

The upper and lower interval limits of the frequencies have been plotted with the purpose of identifying the correct distribution of the variable. Moreover, the variances and means of the limits were calculated. The plots clearly show that the distribution of the interval limits is skewed, but the mean value and the variance have different values. (See appendix) Therefore it can be stated that the limits are not Poisson distributed, which does not necessarily reject the possibility that the frequencies as such are Poisson distributed. The observed distribution of the interval limits is in part a function of the design of the questionnaire, which means that it can only be a crude estimation of the true distribution of frequencies.

It is desirable to recode the data into general specifications, avoiding unnecessary limitation of the interpretation of the results. Therefore, the town districts in Linköping where the respondents live have been translated into four different categories: inner city, residential areas closest to the inner city, suburban residential areas and detached villages. This measure will also have a healthy influence on the number of variables in the data set. The income classes are approximated by the middle of each particular class. The differences in income, before and after the out-of-town establishments, were translated into percentage measures such as +10 per cent, -10 per cent, and +-0. A dummy variable was introduced to test the influence on purchasing frequencies caused by changes in the supply and functions of stores. This variable was given the value 1 for the data describing different variables after the establishment of the out-of-town stores, and the value 0 for the data describing the "before" situation. This means that the observations from a household constitute two cases in the applied data set, with the frequency "before" and "after" as the independent variable.

The Poisson regression analysis

In contrast to e.g. purchased volume, it is not plausible to assume that the purchasing frequency is normally distributed, in view of the fact that the variable only contains integers. The Poisson distribution is well suited for modelling frequencies and has been used extensively in transportation research. (See for example Brännäs, 1987 and Goulias, 1999). Poisson regression and the related negative binomial models have also been used in marketing and shopping research. (Rosenqvist et al, 1989) The Poisson distribution only contains integers, is limited to positive values and is, in contrast to the normal distribution, skewed. The mathematical form of the Poisson distribution is described in (2):

$$(2)\ \text{Prob}(Y{=}y) = \frac{e^{-\lambda_i}\lambda^y}{y!},\ y = 0, 1, 2, \ldots$$

The following diagram shows 500 Poisson distributed randomised numbers with the expected value 5. As can be seen in the diagram the distribution is skewed to the right, with a rather long tail on the right. Some households prefer to buy groceries in small amounts with a high frequency, and some prefer lower frequencies. But all households have to purchase groceries with some minimum frequency higher than zero, which indicates that the Poisson distribution resembles purchasing frequencies rather well.

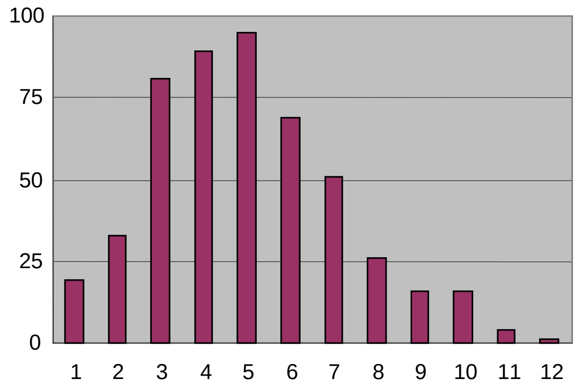


Figure 1: *A Simulated Poisson Distribution with the Expected Value 5*

The purpose of the Poisson regression model is, of course, to explain the expected value with some relevant explanatory variables. It is common to assume that the function of the expected value is exponential, but in principle any form of function can be used. We apply an exponential function with the base *e* in this case, which is a popular approach because the logarithm of the likelihood is then easy to deal with. The expected value is expressed as in (3):

$$(3)\quad \lambda(\mathbf{x}_i, \boldsymbol{\beta}) = e^{\boldsymbol{\beta}'\mathbf{x}}$$

One assumption, which follows from the Poisson regression approach and the basic properties of the distribution, is that the variance and mean of the dependent variable are identical. It should in general be tested whether this is true or false in the data set in question. Examples of relevant tests can be found in Greene (1997). If the assumption is proven to be false, the possibility remains to use negative binomial models instead of the Poisson regression. But calculation of the residuals from a grouped data set is an advanced statistical problem that we have left aside in this paper and refer to future research.

The likelihood of the Poisson distribution is described in (4):

$$(4)\quad \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[-\lambda(\mathbf{x}_i, \boldsymbol{\beta}) + y_i \boldsymbol{\beta}'\mathbf{x}_i - \ln y_i! \right]$$

The frequency variable, i.e. purchasing frequency, is divided into different groups as described above, where every group contains one or more of the possible values. This is what is usually called limited dependent data, a relatively common phenomenon in micro econometrics. (Brännäs, 1991, Greene, 1997, Maddala, 1983) The ordinary likelihood function with unique values must therefore be modified. What we want to know is the probability that y is in the interval $[p,q]$, given the expected value λ. In order to deal with this characteristic of the data set, a modified likelihood function is used in the analysis following Lerman et al (1980). The shopping frequency for a household is the sum of the frequencies of shopping in three different stores. Every frequency value is measured as one of seven possible intervals, which makes 343 potential groups (7^3=343). The grouped log-likelihood is expressed in (5).

$$(5)\quad \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{I} \left[\ln \sum_{n=p_i}^{q_i} \lambda(\mathbf{x}_i, \boldsymbol{\beta})^n / n! - \lambda(\mathbf{x}_i, \boldsymbol{\beta}) \right]$$

The covariance matrix for the estimated parameters is obtained as the negative of the inverted Hessian, i.e. the second derivatives matrix. (Lerman, 1980)

Since $e^{\beta(x+y)} = e^{\beta x} e^{\beta y}$, an increase of x_i by one unit will cause a relative change in the expected value by a factor e^{β_i}. The expected value is not linear and the absolute influence on the expected value due to a change in one variable, i.e. the partial derivative, is therefore dependent on other variables and coefficients in the specified model.

The model was implemented in Microsoft Excel and the estimates of the parameters were calculated with the maximum likelihood method. The maximisation problem was solved with the Problem Solver, an optimisation routine in Excel. Since the function is globally concave, which of course implies the existence of a unique maximum, it is rather simple to find the estimates of the parameters that maximise the likelihood. (Lerman et al, 1980) The computing software Matlab was used for the computation of the variance and covariance matrix.

Results

"Backward elimination" was the guiding principle for the specification of the model. The first step was to use all the potential explanatory variables in the data set, and then eliminate the variables with no significant influence on the shopping frequency one at a time. One small deviation from this principle was made when the variable "driving licence" was eliminated. The reason for this was the strong correlation between this variable and the variable "access to a private car". The latter variable seems to be a better choice from a logical point of view. This switch has of course a negligible effect on the outcome of the analysis, due to the strong correlation between these two variables. It was found during the estimation of the model with different specifications, that reducing the explanatory variables had only a small influence on the coefficients for the remaining variables and none of the coefficients alter signs. These facts indicate the likelihood of stable estimates.

The result of the analysis is summarised in the following table:

Table 1*: The estimated coefficients with goodness of fit*

	Coefficients	Marginal influence	Standard deviation	t-value
Constant	2,1308		0,0477	44,7158
Before/after	0,0421	1,0430	0,0121	3,4704
Age	0,0008	1,0008	0,0004	2,2216
Car access	-0,0633	0,9387	0,0163	-3,8697
Detached village	-0,1071	0,8985	0,0228	-4,6992
Household members	0,0676	1,0699	0,0092	7,3785
Children	-0,0700	0,9324	0,0067	-10,5060
Ln volume	0,0493	1,0505	0,0064	7,7275

			The share of correctly fitted values			
Likelihood	**LR-test statistic**	**p-value**	**+-0**	**+-1**	**+-2**	**+-3**
-4048,96	189,2426	0,0000	36,09	46,33	57,14	67,06

The likelihood ratio test statistic is compared with a χ^2-table with degrees of freedom equal to the number of restrictions imposed on the restricted model, which in this case means the number of explanatory variables in the unrestricted model. The test statistic is large enough for us to conclude that the model is significant.

The average number of shopping trips, i.e. frequency per month, is influenced in the following way by the explanatory variables:

- Increases by four percent as a result of the new out-of-town stores
- Increases marginally with the age of the respondent
- Decreases by six per cent if the household has a car
- Decreases by ten per cent if the household lives in a detached village
- Increases generally with the numbers of household members, but decreases with the number of children
- Increases with the purchased volume of groceries (Ln volume)

The volume variable is a logarithm because the normal equivalent could not be estimated with a significant coefficient. Using the logarithm is of course an ad hoc method without any formal linkages to economic or statistical theory. The logarithm of the expected value is an additive function of the natural logarithm of the volume variable, as a result of the specification of the model. It is also clear that the model as such is statistically significant, in view of the fact that 36 percent of the estimated frequencies coincide with the true interval. Allowing some margins further increases the share of correctly fitted values. Thus it can be stated that the true observations are scattered in close proximity to the estimated regression line. These conditions seem to imply that specification of the model is correct. It is of course unsatisfactory that tests for overdispersion have not been conducted so far. Therefore, it is not clear if the purchasing frequency really is an example of a Poisson distributed variable, or if it would be better to use a negative exponential model. To find and perform such tests on grouped data is an important task for future research.

We have in any case estimated a significant model with interesting results. The approach applied in this paper not only generates the signs of how different explanatory variables influence the shopping frequencies and shopping trips, it also gives information about the relevant magnitudes of these influences. Therefore it seems that the Poisson regression method is a much better choice than the standard procedure with cross tables.

Conclusions and questions for further research

We have identified a set of explanatory variables that influence the frequency of purchasing trips and we have quantified this influence. The results are preliminary and need to be integrated in a more comprehensive framework. We are fully aware of the fact that this paper lacks a thorough discussion of the underlying theory, assumptions and causal relations between factors and variables. We intend to address these, and other, questions in forthcoming papers. The research presented here is part of a broader research programme, where we are trying to appraise the economic efficiency of different grocery distribution systems in urban settings. The nucleus of the efficiency problem is to balance economies of scale in the distribution firms against costs associated with households participating in the distribution of groceries. The increased costs incurred by consumers when transporting, handling and storing groceries can counterbalance the lower price levels in the stores.

Our ambition is to construct an empirically based spatial optimisation model, which can be used to define the overall welfare maximum and evaluate different distribution systems. (Haraldsson 2000) Therefore, it is necessary to perform analysis such as the work discussed in this paper to obtain input for the specification of the model. An important factor that supports model analysis in this research field is the lack of empirical data about Internet shopping in grocery retailing. It is impossible to judge, merely on the basis of figures that describe the financial results of the particular firms, market shares etc., the economic efficiency and potential of Internet shopping compared with a regular store based system. Instead we need analytical tools that can integrate the distribution costs in both the formal and informal sector in order to appraise the overall economic and social efficiency of Internet shopping. There is one particular hypothesis that demands immediate treatment: Is Internet shopping in grocery retailing a superior way of distributing goods that is blocked by market failures caused by the present organisation of the economy in formal and informal sectors?

References

Activity-based travel forecasting conference, June 2-5, 1996. Summary, recommendations, and compendium of papers, US Department of Transportation, Washington, 1997

Brännas, K, *Swedish travel frequencies*, Statistical research report 1987-5, University of Umeå, 1987

Brännäs, K, *Limited Dependent Poisson Regression*, Umeå Economic Studies No. 243, University of Umeå 1991

Greene, W H, *Econometric Analysis*, Prentice-Hall, New Jersey, 1997

Haraldsson, M, (in swedish) *Towards a Optimization Model for Grocery Store Location,* Master Thesis, Linköping University, 2000

Lerman, S, Gonzales, S, ”Poisson Regression Analysis Under Alternate Sampling Strategies”, *Transportation Science*. Vol 14, 295-406, 1980

Ma, J, Goulias, G, ”Application of Poisson Regression Models to Activity Frequency Analysis and Prediction”, *Transportation Research Record*, vol 1676, 1999

Maddala, G S, *Limited dependent and qualtitative variables in econometrics*, Cambridge, 1983

Neter, J, Kutner, M H, Nachtsheim, C J, Wasserman, W, *Applied linear statistical models*, Chicago, 1996

Rosenqvist, G, Strandvik, T, *NBDX-The negative binomial model with explanatory variables*, Swedish School of Economics and Business Administration, Working Papers 197, Helsinki, 1989

Svensson, T, (in swedish), *Structural change of food distribution- driving forces and consequences for urban design and environment*, Linköping Studies in Arts and Science, 179, Linköping University, 1998

Appendix

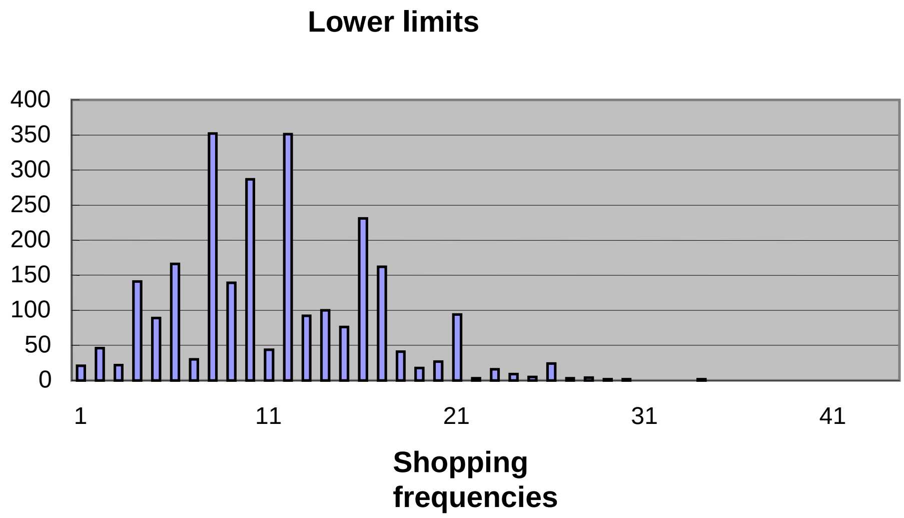
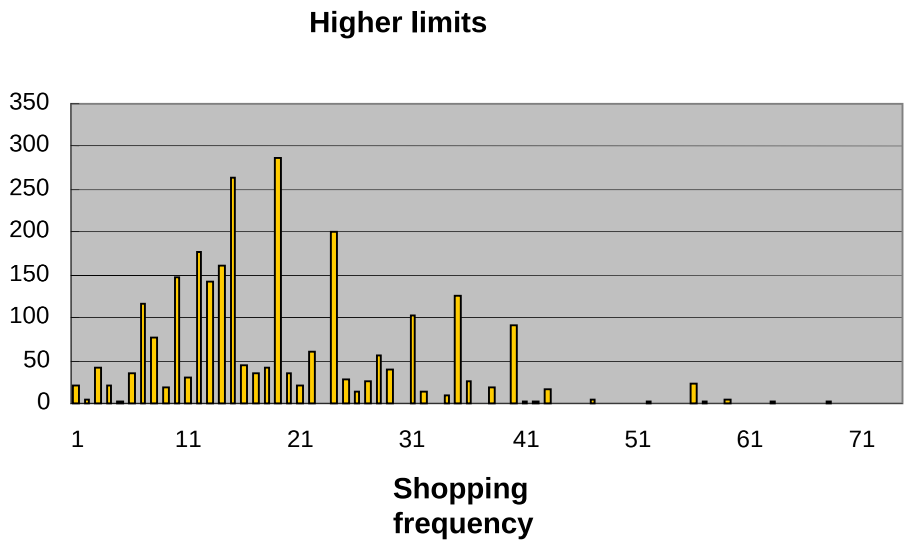


Diagram 1: Shopping frequency intervals, lower limits

Source: Svensson(1998)

Diagram 2: Shopping frequency intervals, higher limits



Source: Svensson(1998)

Table 1: Mean and variance

	Lower limits	Higher limits
Mean	11,3832	19,3251
Var.	26,8346	110,7353

Source: Svensson(1998)