Mònica Casabayó
Dept. Marketing
ESADE, U. Ramon Llull
Av. Pedralbes, 62
08034, Barcelona (Spain)
+34 932806162
casabayo@esade.edu

Núria Agell
Dept. Quantitative Methods
ESADE, U. Ramon Llull
Av. Pedralbes, 62
08034, Barcelona (Spain)
+34 932806162
agell@esade.edu

John Dawson
Marketing Section
The University of Edinburgh
University of Stirling and
ESADE, U. Ramon Llull
08034, Barcelona (Spain)
John.dawson@ed.ac.uk

ABSTRACT

The experiment presented in this paper has used an unsupervised learning technique to forecast online purchasing based on historic in-store data. The methodology is an innovative software tool called LAMDA (Aguilar-Martin and López de Mántaras, 1982; Aguilar-Martin and Piera, 1986; Aguado, 1998) *based on the fuzzy concept of adequacy* (Aguado, 1998; Casabayó et al., 2004). *Assumed the fact that online purchasing is mainly motivated by shopping convenience, the paper describes how this approach is capable to help retailers to forecast the current customers who are going to buy online. From a managerial perspective, a more realistic way to interpret the results to support decision making in marketing has been introduced as it is capable to deal with ambiguous, uncertain and incomplete information.*

Keywords

Customer behaviour, e-commerce, shopping convenience, unsupervised learning, fuzzy learning technique.

Forecasting customer's behaviour in the Spanish grocery industry: Identifying the customers who are going to buy online

Introduction

The appearance of the Internet meant a new challenge for many companies. Particularly in the food retailing sector, it was known that this new technology could generate a considerable change according to the way firms market and distribute their groceries to customers. Therefore, understanding and predicting online-buying behaviour is of major importance for e-commerce website managers (Bucklin and Sismeiro, 2003). Within the research literature, there are several attempts to support and facilitate the achievement of this managerial goal (Shim and Drake, 1990; Breitenbach and Van Doren, 1998; Douthu and Garcia, 1999; Mathwick, Malhotra and Rigdon, 2002; Shim, Eastlick and Lotz, 2000; Ray, 2001; Chiger, 2001; Supphellen and Nysveeb, 2001 Raijas, 2002; Kim, 2004; Bucklin and Sismeiro, 2003; Van den Poel and Buckinx, 2005).

The experiment carried out is based on internal data gathered from the loyalty card and scanner systems of a Spanish supermarket chain.

The main goal of the research problem is to learn from the current customers' specific behaviour to predict their own individual behaviour but in a different situation. Observed historic behavioural data has already proved to be commonly used as an effective predictor (Schmittlein and Peterson, 1994). Particularly in this experiment, observed behavioural data from the physical store is going to be used to forecast online purchasing.

Defining *Shopping Convenience*

A specific database was built for this experiment. 2.063 customers were selected from 19 stores spread across the city of Lleida. In addition, some variables were chosen as predictors for on-line purchasing. The selection of these variables was based on experts' opinion and research publications. Based on that, *Shopping Convenience* was likely to be the main

motivation to buy online.

Convenience is a fuzzy concept which may take different forms and interpretations. Firstly, consumption convenience is referred to all the products and foods which people normally buy when they don't want or have the time to cook them, such as ready meals from supermarkets or take-aways food for restaurants. In general, consumption convenience takes place when consumers are looking for minimising the efforts and time that they need before and after eating the meal. This type of convenience is not considered in the scope of this research but shopping convenience.

In the literature, shopping convenience is not clearly defined. According to Reimers and Clulow (2000), rather than actually defining the concept of convenience, many researchers simply listed its attributes. To the best of our knowledge, Downs's (1961) was the first research contribution who stated that when seeking convenience, the shopper sought to minimise three costs: money, time and energy. Furthermore, and approximately 30 years later, Gehrt and Yale (1993) identify the temporal, spatial and efforts dimensions when related to convenience.

Table 1 A summary of convenience attributes from a literature review

	Attributes of convenience						
Research Studies	Trading Hours	Proximity	Travel time/Access	Internal layout	Parking	Enclosure	Merchandise Variety
Bellenger, Robertson and Greenberg (1997)		*					
Spencer (1978)		*					*
Howell and Rogers (1980)		*	*				
Cymrot, Gelber and Cole (1982)			*		*		
Timmermans, Van der Heidjen and Westerveld (1982)		*		*	*	*	
Bucklin and Gautschi (1983)				*			
Oppewal, Louviewe and Timmermans (1994)			*		*		
Berrell (1995)							*
Kaufman-Scarborough (1996)							*
Bell (1999)	*	*		*	*		

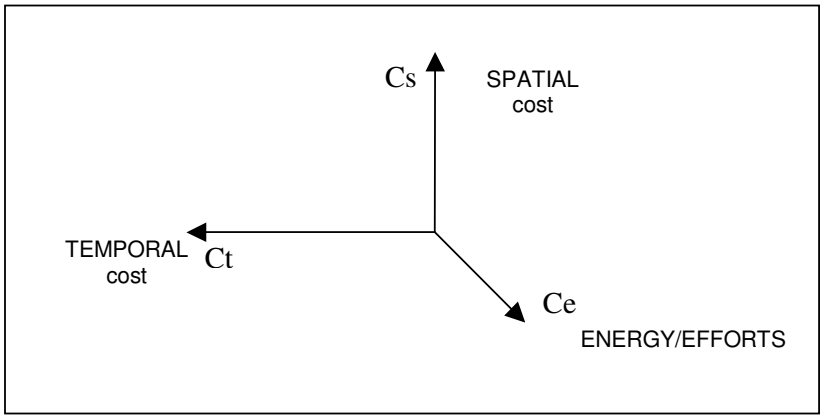
Source: Reimers, V. and Clulow, V. Shopping and Convenience: A model for retail centres. ANZMAC Conference, 2000:8.

As shown in Table 1, seven different categories of convenience in terms of trading hours, proximity, travel time and access, internal layout, parking, enclosure and merchandise variety

have been studied during the last 30 years. All attributes refer to temporal, spatial or energy dimensions, introduced by Gehrt and Yale. However, Timmermans, Van der Heidjen and Westerveld (1982) are the few authors that take the global three dimensional meaning of convenience into account.

On the other hand, as far as the table shows, proximity has been the aspect most directly related to convenience. It is important to note that mentioned publications about shopping convenience were mainly focused on off line retail shopping. However, in this study, the three dimensions of convenience introduced by Gehrt and Yale (1993) are followed as our theoretical approach.

Figure 1 Schematic concept of Shopping Convenience



As shown in Figure 1, the closer location to the centre, the greater is the shopping convenience and the lower is the total cost (Cs+Ct+Ce). A critical part of the experiment is to find out the variables which inherently refer to at least one of shopping convenience's dimensions. It is assumed that the main motivation (in a case when the grocery firm is the same for both channels) that can make the current customer change from SUPSA´s traditional outlet to SUPSA´s online supermarket is shopping convenience. The higher value the customer assigns to shopping convenience (instead of other shopping benefits), the most potential exists to buy online.

In the research literature, publications aiming to compare online shopping and store shopping were also considered when reaffirming the second assumption. Despite the fact that several analyses of the advantages and disadvantages of both retail formats (also called channels) are captured (Strader and Shaw, 1997; Breitenback and Van Doren, 1998; Crawford, 2000;

Degeratu, Rangaswamy and Wu, 2000; Ray, 2001; Burke, 2002) most of the publications directly or indirectly conclude with a common denominator: Although being defined in different ways, convenience is also higher ranked for online shopping than store shopping (Kalakota and Whinston, 1997; Burke, 2002; Dahlén and Lange, 2002).

According to Raijas (2002), convenience is expected for either physical store customers or store website customers. However, the website not only can achieve the traditionally most important factors affecting the physical store choice (low price level, customer service, location, product assortment) but also it is able to avoiding all the inconvenience of grocery shopping (looking for the products, self picking, waiting queues, self delivering .etc). Based on an electronic grocery shopper's survey, Raijas (2002:111) concludes that

> *'the principal benefits for online purchasing are (1) time and effort saving, (2) time and place independence and (3) possible tools for follow-up and planning'.*

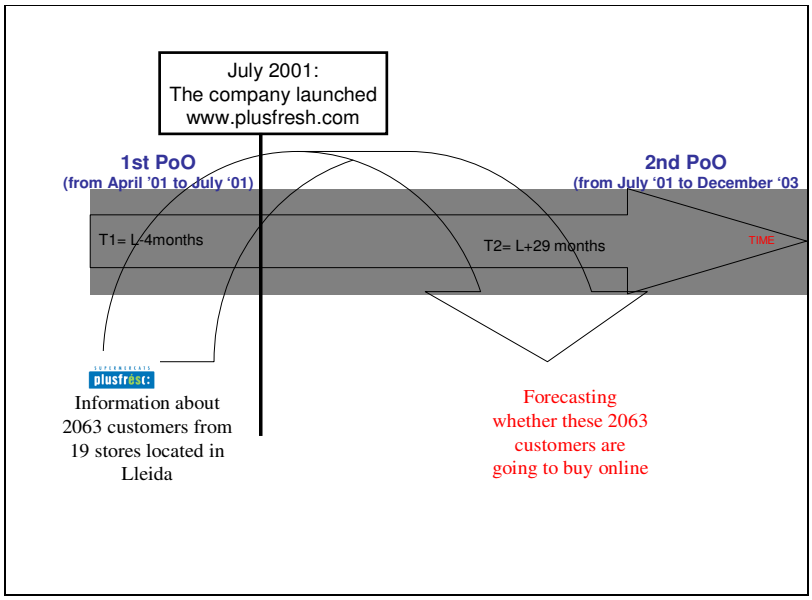
External secondary sources supported the assumptions. For example, according to AECE's (2000; 2003) results, the first Internet users' motivation to buy online is convenience.

Once the main online shopping motivation was determined, the variables that could define shopping convenience were selected. Mention that the information stored in the internal database was mainly behavioural and socio- demographic. There was no data corresponding to either customers' Internet perceptions or customers' online purchasing intentions, as data all come from the loyalty card programme and scanner systems.

Selected predictors for *'Shopping Convenience'*

In order to identify shopping convenience, 28 indicators were initially selected for the experiment. Noting that by merging different categories of data, the predictive power of the modelling exercise is maximised (Montgomery, 2001); the predictors were split into two main categories. The first was termed 'socio-demographic details'. The second one grouped the observed 'behavioural in-store data'. It is important to note that all these indicators correspond to the first period of observation (PoO).

Figure 2 Determination of the Periods of Observation (PoO)



The first one (T_1) comprises the four months prior to the launch of the company's website. The second period is longer. The second period (T_2) is the interval between the moment the company launched the website (July 2001) and the time when there were sufficient customers to test the efficiency of the forecasts. It is important to note that initially, it was thought to take the same period of time (4 months) but the decision to increase the period of observation was taken due to the low Internet penetration ratio.

Socio-demographic details

As previously mentioned, customer demographics have been extensively applied to explain and forecast online purchasing. Consequently, 6 demographic predictors available in the internal database were originally selected. The 6 variables include both the information related to the cardholder subscriber and his/her household. As it is listed in Table 2, the variables mainly focused on the individual characteristics are *customer code, age, gender, employment status*. Mention that *customer code* variable is used to identify each customer (or household). Then, although it is not used for the learning process itself, it is the key data for comparing and analysing the results in the final stage.

Directly related to the household are V5 (address) and V6 (email). There was no information available related to the household income. Then, the address was chosen as it was likely to provide information about the area where the customer lives (rich, medium, poor area of the

city).Variable 6 was also interesting for the purpose of the experiment. Having *an email* was assumed to be positively related to having Internet connection and frequent accessibility. Hence, it seemed to be a highly significant variable.

Table 2 Socio demographic predictor variables selected for the experiment

N	Variable name	Description
Information related to the individual Cardholder		
1	Customer code	The loyalty card code which allows to identify the customer
2	Employment Status	There is a classification between: housewife, retired, unemployed, employed, employee, self-employed
3	Gender	Women/Men
4	Age	Date of Birth
5	Address	This variable gives information about the area where the customer lives,.
6	Email	Yes or No. The answer indicates whether he/she has Internet access and frequency of access.

Despite the fact that race and language have been also considered interesting predictors (Padmanabhan, Zheng and Kimbrough, 2001), these were not applicable in Lleida, where the major spoken languages are Catalan and Spanish. Moreover, despite the fact that immigration is slightly increasing, SUPSA´s customers are dominantly Spanish.

Observed behavioural in-store aspects

As is listed in Table 3, the *Observed buying behaviour data* captured from purchases in traditional stores joins 21 predictors classified according to the 3 shopping convenience dimensions: *Temporal, Energy/effort and Spatial* categories. The objective of this selection is to establish a concrete description of the shopping convenience concept for the experiment.

Despite the fact that almost all the indicators could be located in more than one category, possible overlapping is ignored. Accordingly, each predictor has been only located to one single dimension.

Table 3 Classification of the observed behavioural in-store aspects.

DIMENSIONS OF CONVENIENCE (Gerht and Yale, 1993)	Number of predictors
TEMPORAL DIMENSION	5
ENERGY EFFORT DIMENSION	9
SPATIAL DIMENSION	7
TOTAL CONVENIENCE CONCEPT	**21**

As shown in Table 4, the *temporal dimension* is represented by 5 predictors. When individuals experience high levels of time scarcity, they are likely to have certain ways of thinking about and using time (Kaufman-Scarborough and Lindquist, 2002). Therefore, a selection of specific indicators able to deduce degrees of time scarcity has been chosen for the experiment such as % delivered purchases after seven p.m. Indicators more related to frequency are collected as well. According to Raijas (2002), online shopping frequency is lower than in a conventional grocery store. Based on his statements, variables 7, 8, 9, 10 and 11 were selected.

Table 4 Observed behavioural in-store data predictors selected for the experiment: Selection of the convenience's temporal dimension predictors

N	**Variable name**	**Description**
TEMPORAL DIMENSION		
7	% delivered purchases after 7 p.m.	Percentage of monetary value referred to the deliveries after 7 pm (according to the total amount spend during T1)
8	Mean No. of days per week on which purchases are made	Average number of trips to the shop by week
9	% purchases made on Saturdays (number of customer trips)	Percentage of monetary value made on Saturday (from the total shopping trips made during T1)
10	% purchases made on Saturdays (amount)	Percentage of monetary value spend on Saturdays (from the Total Purchase of T1)
11	% of purchases made from Monday to Wednesday (amount)	Percentage of monetary value spend from Monday to Wednesday (from the Total Purchase of T1)

In reference to *effort/energy dimension*, 9 predictors were determined (See Table 5). According to Raijas´s (2002) contributions, despite the fact that online grocery shoppers tend to buy the same products as in a conventional store, they tend to concentrate purchases of dry products and beverages. On the one hand, *fresh products* are bought less online. On the other hand, avoiding the picking and handling were also relevant points when choosing online grocery shopping. Supporting this, variables from 12 to 19 were selected. The number of outlets where the customer purchase also informs about an additional effort, therefore whether the customer buys in more than one is captured in this variable (V20).

Table 5 Observed behavioural in-store data predictors selected for the experiment: Selection of the convenience's energy/efforts dimension predictors

ENERGY EFFORTS DIMENSION		
12	% purchase of fresh produce /Total purchases	Percentage of monetary value spend fresh produce from the total items purchased
13	% meat purchases at self-service counter / Total purchases	Percentage of monetary value spend on meat at self service from the total items purchased
14	% of purchases made up by special offers/ Total packed product	Percentage of monetary value spend on special offers from the total packed items purchased
15	% delivered purchases (amount)	Percentage of monetary value delivered at home from the total amount spend on T1
16	% delivered purchases (nº of customers trips	Percentage of delivered purchases at home from number of customer trips to the shop during T1
17	Was auto-scanning used?	Some SUPSA's stores have auto-scanning service. This data informs whether it was used by the customer or not.
18	Means of transport	Based on the address information this variable indicates the distance to a store. This variable informs whether the customer comes by food, walking or by car.
19	% of coupons and discounts redemption	From all the coupons and discounts launched by the company, this variable measure the % that the customer uses them.
20	Number of outlets where customers shop	Number of SUPSA's stores used by each customer

Particularly, an explanation of *was scanning used* (V17) and *meat purchases at self service* (V13) is required as they are special features of the company. Auto-scanning is not available in every store. Customers are given an easy-to-use device which helps them to save time when checking out. Customers do not need to wait to have their items scanned, because they have already scanned their items while they were walking and picking them from the aisle. Information corresponding to *Meat purchase at self service* is captured by the scanner systems every time the customer takes it directly to the meat shelf instead of waiting for his turn in the Butchery inside the supermarket store.

Special sales and coupons was one of the most relevant attributes of performance at web stores (Chiang, Zhang and Zhou, 2004). Therefore, variable 14 *(% of purchases made up by special offers and % of coupons and discounts redemption*) were also included for defining the effort dimension of convenience. *Means of transport* (V18) was resulted from the transformation of the variable *address*. This resulted variable was split between 3 categories, which include 'by foot', 'walking' and 'by car'.

Referring to the *spatial dimension* of convenience, Table 6 shows the variables that correspond to the location and distribution of the products in the store. In particular, 4 indicators were selected to describe this type of convenience category (V24, V25, V26 and V27).

Table 6 Observed behavioural in-store data predictors selected for the experiment: Selection of the convenience's spatial dimension predictors

SPATIAL DIMENSION		
21	Average spent per item (Total T1)	The money spent by item purchased during T1 divided by the number of items
22	Average purchase (per total trips)	From all the purchases that the customer realises during this period, this variable shows the average of purchases
23	Total Purchases	Total monetary amount of spending during T1
24	Size of the outlet	SUPSA's classify their stores into big, medium or small.
25	No. of different articles purchased in the period	Number of different items purchased during T1
26	No. of departments where no purchases were made	Number of departments were NO item was bought during T1
27	Number of TOTAL items purchased	Number of references bought during T1

Furthermore, Raijas (2002) suggested that the average amount spent in electronic grocery shopping was generally higher than the amount spent in the store. Based on this suggestion, *Average of items purchased*, *Average purchase with the company* and *Total Purchases* were taken into account for the experiment and included in the spatial dimension (See Table 6).

Predicted variable: 'Online purchasing behaviour'

No previous SUPSA based historical data corresponding to online purchasing existed, so an assumption was required before the implementation of LAMDA´s approach.

- Assumption: The customers who previously bought by distance selling (fax, telephone, email) are considered the strong potential Internet buyers.

SUPSA had information on the customers who had ordered their purchases either by fax, email or telephone. 78 out of 2.063 had demonstrated attributes that showed them to be more interested in shopping convenience than in other store benefits. These 78 customers were noted and labelled as distance-buying individuals. This variable is not considered in the learning process. It is only going to be used as a criterion to decide the most relevant classification from the wide range of results provided by the unsupervised LAMDA algorithm. V28 is the predictive variable, crucial when forecasting the online customers.

Table 7 Determination of the predictive variable

N	Variable name	Description
Predictive Variable		
28	Has the customer ever purchased by distance selling?	Despite the fact that the company did not have a Website, their customers had been able to order their shopping basket by telephone, fax or email.

Validating corpus

The variables presented in this section correspond to the second period of observation (T2). In December 2003, the company checked the customers who bought, at least once at www.plusfresh.com (See Table 8).

Variable 8 Variables related to the real online purchasing customer

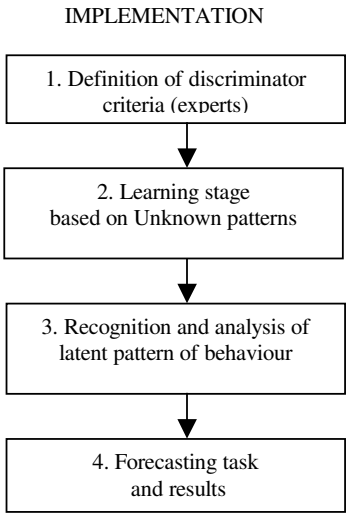
N	Variable name	Description
Real online behaviours (Variables related to PoO T2)		
29	Has the customer ever bought at www.plusfresh.com?	Yes or No (from July 2001 to December 2003)

V29 informs whether the customer has bought at least once at plusfresh.com since it was launched in July 2001.

Implementation of the unsupervised forecasting model

LAMDA (Aguilar-Martin and Piera, 1986; Piera and Aguilar-Martin, 1991; Aguado, 1998) is a classification method based on hybrid connectives (Aguilar-Martin and Piera, 1986). These hybrid connectives allow the several forms of partial information, resulting from the relationship between the individual to each variable (marginal adequacy degree) to be turned into a simple result which assigns each individual to an appropriate existing segment. Then, there are as many marginal adequacy degrees as variables (predictors) used in the experiment. Moreover, there are as many global adequacy degrees (GAD) as the number of existing segments considered in the experiment. LAMDA gives the possibility of forecasting based on supervised learning and unsupervised learning. Figure 3 illustrates the implementation stages of LAMDA's unsupervised forecasting model.

Figure 3 LAMDA´s unsupervised forecasting model



Once the selection and analysis of the predictors are finished, the implementation of unsupervised forecasting model starts. Due to the large number of classifications which will result from the learning and recognition stages, the first step is focused on defining the discriminator criteria which will help to choose between them. The learning task is performed in the second stage. Afterwards, the recognition of latent patterns of behaviour and its analysis is developed. The fourth step consisted of the forecasting task. Finally the results were collected.

Defining the discriminator criteria

Unlike the supervised model, measuring and labelling the existing patterns of behaviour was not required as no previous pattern of behaviour existed. The unsupervised learning algorithm first learns from customers' information, recognises afterwards latent patterns of behaviour and suggests a set of possible classifications to be analysed. In order to select between the vast options provided by the model, some criteria to discriminate between the possible classifications has to be decided at this point. A meeting with marketing personnel from SUPSA was held. The purpose of the meeting was to know the decision process and the criteria that experts would use to identify and forecast 'online shopping behaviours' without disposing of any particular market research survey. Interesting conclusions resulted from the meeting and 3 discriminator criteria were established. The selection of the most appropriate classification was carried out by means of these next criteria:

a) The classification has to follow the grouping rule

The grouping rule is measured by the concentration of the distance-selling individuals in one or more segments of the classification. The customers who had purchased by distance selling would be used to determine the customers who would buy online. Then, the classification that presented at least one segment with a high concentration of distance buyers would be considered. Particularly, if one of the segments of the classification joined more than 25% of the -buying individuals (that means more than 21 distance buyers in absolute numbers), the classification was marked as 'interesting'.

b) The classification has to be manageable

Apart from the grouping rule, the classification should be manageable. The unsupervised learning algorithm suggests a wide set of classifications based on the latent pattern of behaviour that the algorithm recognised. Then, each classification is likely to be composed by different numbers of segments. From management point of view, it was decided that a classification which showed more than 5 segments was considered not manageable as it is difficult to be interpreted.

c) The classification has to be balanced

Classifications resulted from the recognition stage are also likely to present unbalanced segments. A classification is unbalanced when one of its segments groups has more than the 80% of the total individuals (2.063). In that case, there is virtually just one segment, and when interpreted from a marketing point of view it is not useful as little discrimination between individuals is provided.

In addition to these 3 criteria, there are also 2 conditions inherently related to LAMDA which have to be considered as well. These are the following:

d) The classification has to be stable

A stable classification takes place if at the recognition stage, all individuals are reassigned in the same segment. When applying the unsupervised learning, the individuals who are located firstly to some segments have a higher weight to the segment than the last ones. Once the algorithm learns, the individuals tend to be located in the same segment. When the learning process assigns repeatedly the same individuals in the same segments, the classification is stable.

e) The classification has to be unique

The classification is likely to be proposed more than once. Different combination between LAMDA's capabilities may suggest exact patterns of behaviour. All the repetitive classifications are directly removed by this criterion.

Having explained the discriminator criteria, it is important to note that just the classifications which fulfil the five criteria will be selected as promising.

1. Learning stage

The unsupervised learning process takes place when the different types of hybrid connectives provided by LAMDA are combined automatically with a specific level of tolerance. This is a trial and error approach based on an iterative process. Then, several combinations of three elements which include a specific number of iterations, an applied fuzzy connective and an established level of tolerance are tested. The learning process is considered to be finished when either the classifications are stable or a pre-determined number of iterations are carried out. In this experiment, 10 is the pre-determined number of iterations.

2. Analysis of Recognised latent pattern of behaviour

Using the unsupervised learning capabilities of LAMDA algorithm, 945 classifications were obtained. Table 10 shows the combinations applied for the experiment. Based on the same number of iterations and with an automated tolerance selected, minmax algorithm was able to recognise 815 different latent patterns of behaviours while probabilistic algorithm just recognised 7.

Table 10 Classifications resulted from LAMDA's unsupervised learning capabilities

Fuzzy connective	Number of iterations	Tolerance	Number of classifications resulted
MINMAX	10	Automatic	815
FRANK	10	Automatic	123
PROBABILISTIC	10	Automatic	7
TOTAL			**945**

The number of classifications resulted from each algorithm is not important. What really matters is how many classifications accomplish the discriminator conditions. The following table explains the finalist classifications according to the unique and stable criterion:

Table 11 Finalist classifications based on unique and stable criteria

		UNIQUE CRITERION	STABLE CRITERION
Fuzzy connective[1]	Number of unique classifications obtained	Number of unique classifications	Nº of Stable classifications from the total obtained
MINMAX	815	608	29
FRANK	123	98	3
PROBABILISTIC	7	7	1
TOTAL	**945**	**713**	**33**

Despite the fact that initially, 945 classifications were obtained, just 713 were not repetitive classifications. Noting the initial 815 classifications suggested by minmax algorithm that were reduced to 608. Moreover, when analysing the stability of these 713 classifications, a dramatic reduction of the number of possible finalists took place. Just 33 classifications out of 713 accomplished the stable criteria as well, 29 came from Minmax, 3 from Frank and just one from Probabilistic. The next step was to analyse the 33 classifications according to the rest of the criteria. As it is shown in Table 12, all the classifications recognised by FRANK and PROBABILISITIC presented less than 5 segments. However, just 18 classifications out of the 29 resulted from MINMAX satisfied this managerial criterion.

Table 12 Finalist classifications based on the managerial criterion

Nº SEGMENTS	2	3	4	5	6	7	+ 7	FULFILED MANAGEABLE CRITERION
MINMAX	7	0	2	9	2	3	6	18
FRANK	1	1	0	1	0	0	0	3
PROBABILISTIC	1	0	0	0	0	0	0	1
TOTAL								22

Then, the balance criterion was analysed as well. Firstly, the 18 classifications from minmax were analysed. None satisfied the balanced criterion. As far as the 3 FRANK classifications are concerned, one was removed as more than 80% of the total individuals were grouped in

[1] The number of unique and stable classifications obtained by Lukasiewicz was 0.

the same segment. Table 13 shows the only two classifications that fulfilled all the criteria.

Table 13 Finalists classifications based on balanced criteria

		BALANCED CRITERION					FULFILED CRITERIA
Classification	Tolerance	% of individuals grouped in each segment					The 4th first criterion are fulfilled
2. FRANK	0.443	54%	21%	17%	14%	3%	YES
3. FRANK	0.454	66%	30%	4%			YES

Finally, the last criterion (grouping criterion) was checked in order to know whether the classification was suitable to forecast online purchases. Both, FRANK 0.443 and FRANK 0.454 were analysed:

Table 14 Composition of classification 2, Frank 0.443

2. FRANK 0.443			Total segment	% distance buyers
Segment 1	*Distance buyers*	*22*	928	**28%**
	Others	906		
Segment 2	*Distance buyers*	*20*	427	**26%**
	Others	407		
Segment 3	*Distance buyers*	*8*	357	10%
	Others	349		
Segment 4	*Distance buyers*	*25*	285	**32%**
	Others	260		
Segment 5	*Distance buyers*	*3*	66	4%
	Others	63		
TOTAL	*Distance buyers*	*78*	2063	100%
	Others	1985		

As illustrated in Table 14, segment 1, segment 2 and segment 3 present a concentration ratio of distance buyers of 28%, 26% and 32% correspondingly. The grouping principle is perfectly fulfilled.

Table 15 Composition of classification 3, Frank 0.454

3. FRANK 0.454			Total segment	% distance buyers
Segment 1	*Distance buyers*	*38*	1353	**49%**
	Others	1315		
Segment 2	*Distance buyers*	*36*	625	**46%**
	Others	589		
Segment 3	*Distance buyers*	*4*	85	5%
	Others	81		
TOTAL	*Distance buyers*	*78*	2063	100%
	Others	1985		

Also segment 1 and segment 2 from classification 3 (See Table 15) accomplish the grouping criterion as both of them concentrate more than 25% of the total distance buyers.

3. Forecasting task

It is important to remark that all the individuals placed in the same segment shares the same pattern of behaviour, although this pattern is unknown. For the forecasting task, variable V28 was mainly used. Since the experiment had no ideal partition to conduct a comparison, the 78 clients who had engaged in a remote buying act by e-mail or fax, called distance-buying customers were assumed to be almost certain customers for web purchase. Accordingly, customers who were located in the same segment as the distance buyers were considered potential online customers as well. Based on that, from classification 2 (See Table 13), segments 1, 2 and 4 would be selected as the online buyers. That means that the 928, 427 and 285 customers respectively would be considered potential online buyers. From classification 3 (See Table.14), the customers located in segment 1 and 2, which are 1353 and 625 respectively would also be the potential online buyers.

From the two final classifications, a selection of just one was required. However, there were no objective grounds for making a final selection between the two because they did not have the same number of segments. To solve this apparent tie-break, we tried to reduce the 5 segments from Classification 2 to 3 segments so a final comparison was then enabled. The reduction of the number of segments is carried out by the intervention process.

In LAMDA, each individual belongs to all the segments to a greater or lesser extent. Although each individual is finally allocated to the segment to which it presents the greater

maximum GAD, it still belongs to the rest of the segments with a specific GAD to each segment. Therefore, each individual is thus assigned an adequacy rating (vector). Looking at the individual vector, it is possible to observe what happens if we multiply one of these adequacy ratings by a corrective parameter. Table 16 shows the corrective parameter and how it causes the vector to change, forcing the individual into another segment.

Table 16 Intervention process applied in classification 2, Frank 0. 443

Segment 1	**0.98**	**1**	**1.02**	**1.04**	**1.06**	**1.08**	**1.1**	**1.12**	**1.14**	**1.16**	**1.18**
1	875	928	1053	1176	1226	1249	1288	1347	1426	1518	1595
2	465	423	356	304	285	282	276	262	246	226	208
3	362	357	348	313	310	293	276	255	220	184	151
4	295	289	242	207	181	178	166	143	118	86	63
5	66	66	64	63	61	61	57	56	53	49	46
Segment 2	**0.98**	**1**	**1.02**	**1.04**	**1.06**	**1.08**	**1.1**	**1.12**	**1.14**	**1.16**	**1.18**
1	975	928	882	862	857	848	836	812	764	699	622
2	315	423	517	620	657	685	728	793	893	1032	1147
3	358	357	353	315	310	293	278	259	226	193	176
4	348	289	248	204	178	176	164	143	127	90	72
5	67	66	63	62	61	61	57	56	53	49	46
Segment 3	**0.98**	**1**	**1.02**	**1.04**	**1.06**	**1.08**	**1.1**	**1.12**	**1.14**	**1.16**	**1.18**
1	928	928	923	912	896	846	762	653	569	469	356
2	425	423	421	383	363	347	330	295	265	231	181
3	310	357	400	484	536	605	731	904	1037	1218	1406
4	334	289	254	219	204	201	181	153	137	96	74
5	66	66	65	65	64	64	59	58	55	49	46
Segment 4	**0.98**	**1**	**1.,02**	**1.04**	**1.06**	**1.08**	**1.1**	**1.12**	**1.14**	**1.16**	**1.18**
1	954	928	915	890	854	832	792	724	633	560	497
2	442	423	349	293	275	260	229	207	182	140	90
3	375	357	310	299	270	226	185	162	146	124	94
4	225	289	425	519	606	691	807	926	1063	1207	1357
5	67	66	64	62	58	54	50	44	39	32	25
Segment 5	**0.98**	**1**	**1.02**	**1.04**	**1.06**	**1.08**	**1.1**	**1.12**	**1.14**	**1.16**	**1.18**
1	928	928	927	922	917	908	896	871	821	754	674
2	425	423	418	384	365	361	352	339	322	299	277
3	357	357	353	316	310	293	278	259	226	194	176
4	290	289	268	230	205	203	189	167	147	108	84
5	63	66	97	211	266	298	348	427	547	708	852

For instance, when multiplying all the 2.063 GADs in segment 1 for an increasing corrector parameter, it is shown that the customers who initially belonged to segment 2 and segment 4 moved to segment 1 more quickly than the rest. For instance, when multiplying the GADs in segment 1 for a 1.02 corrector parameter, the 423 customers in segment 2 are reduced to 356 and the 289 customers in segment 4 are reduced to 242. These 67 and 47 customers respectively moved to segment 1. The intervention process is applied to each segment from

classification 2. The objective is to observe whether individuals tend to move to other segments when a corrector factor is applied in each segment. Graphically, Figure 6.7 illustrates the intervention process applied in segment 1. As mentioned, all the individuals have a GAD in segment 1. However, just the customers with the maximum GAD to segment 1 are located in it. When all the 2063 GADs in segment 1 are multiplied by the corrector parameter, the individuals which initially had a maximum GAD in segment 2 and segment 4 tends to change to segment 1.

Figure 3 Results from the application of intervention process in segment 1

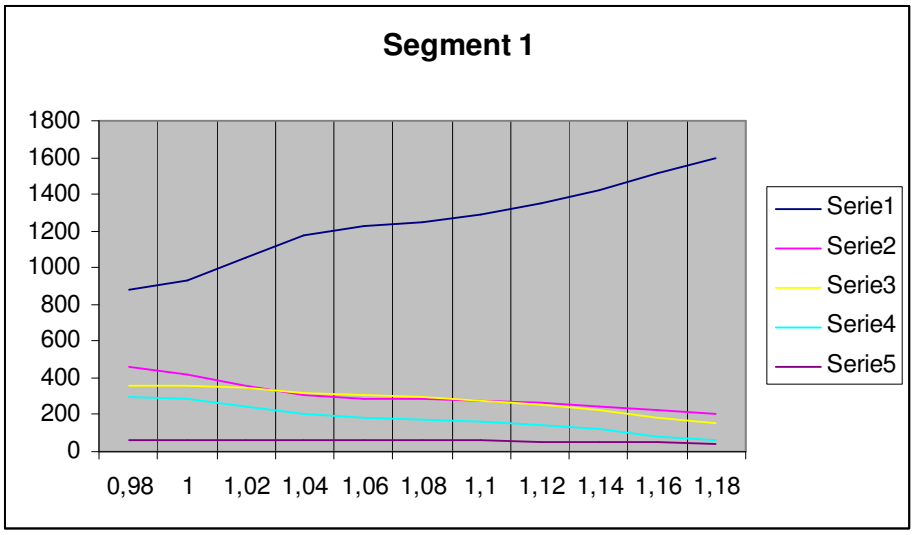
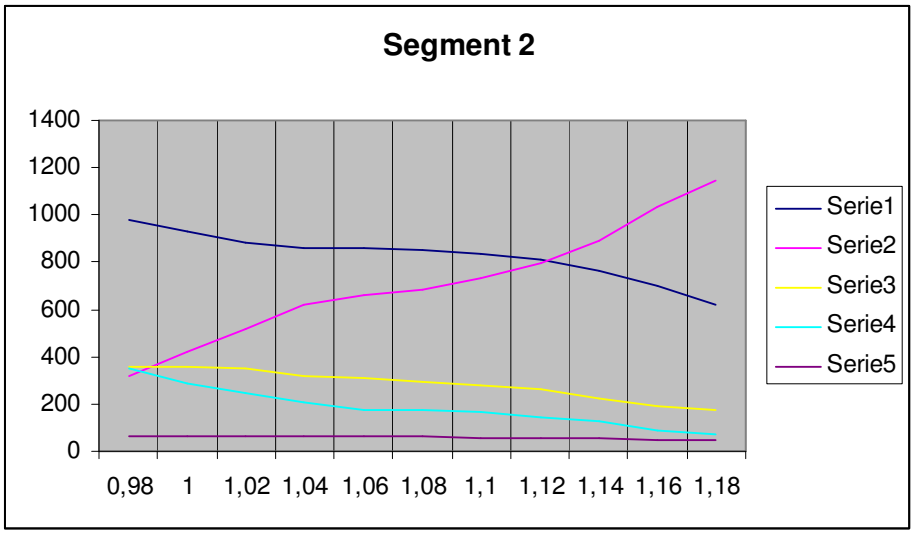


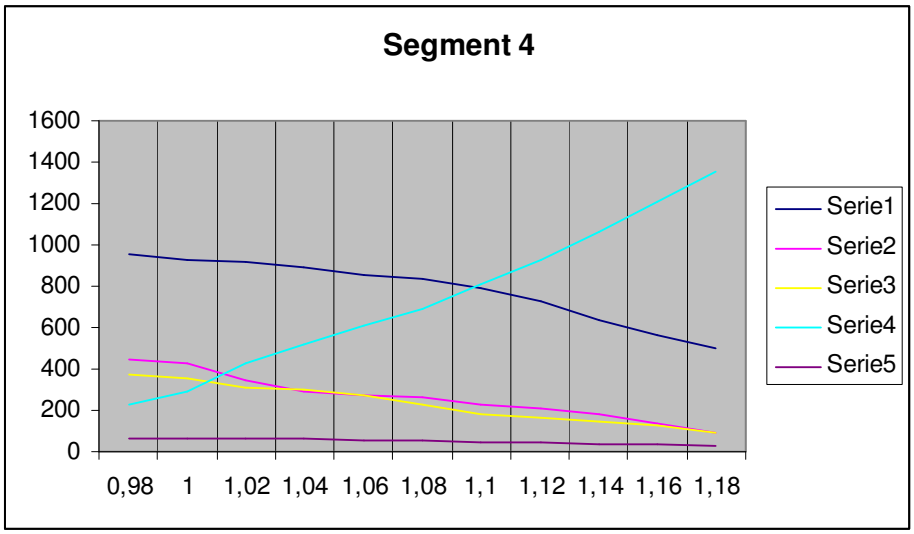
Figure 4 illustrates the intervention process results when the corrector parameter is applied to segment 2. It is evident that when forcing segment 2, their customers tends to move to segment 1 and 4.

Figure 4 Results from the application of intervention process in segment 2



It will be seen that proportionally increasing the adequacy of individuals vis-à-vis segment 2 quickly empties segment 1 and segment 4. When repeating the procedure with segment 4, the following changes occur (Figure 5).

Figure 5 Results from the application of intervention process in segment 4



Having applied the intervention process, it is seen that the segments 1, 2 and 4 are merged in one same segment. Then the 5 segments of the classifications have been reduced to 3 segments to permit comparison with classification Frank 0.454.

Table 18 Classification 2 after the intervention process

3. FRANK 0.443			Total segment	% distance buyers/total distance buyers
Segment 1'= 1+2+4	Distance buyers	67	1640	**85,90%**
	Others	1573		
Segment 2'=3	Distance buyers	8	357	**10,26%**
	Others	349		
Segment 3'=5	Distance buyers	3	66	**3,85%**
	Others	63		
TOTAL	Distance buyers	78	2063	
	Others	1985		100%

The new classification presents 3 segments. The resulted segment 1' groups the majority of distance buyers which represents the 85.9% of the total. 67 distance buyers is the result of joining the number of distance buyers initially located in segments 1, 2 and 4 before the intervention process.

When comparing the number of distance buyers in each segment from classification 2 (See Table 8) with the results from classification 3 (Frank 0.454), we can see that the segment which joins the maximum number of distance buyers is segment 1'. Then, classifications Frank 0.443 is chosen.

Validation

At this stage, the information corresponding to PoO T2 was used. The main goal was to assess LAMDA´s predictions. Forecasts made by the unsupervised LAMDA algorithm were then tested with the information provided by V29, which captured whether the customer had bought online at least once since July 2001 until December 2003. As previously mentioned, a first validity was done in January 2002, but the real online customers were just 93. Therefore, the PoO T2 was extended to December 2003. At that time, online customers had just increased by 10 new online customers.

It must be remembered that, the results were compared and analysed from a marketing standpoint. The main goal was to predict the customers who were going to buy online. As

Table 19 shows, there was the possibility to make two mistakes, but according to our goal the main error was not to identify a real online customer.

Table 19 Interpretation of results based on a marketing standpoint

		LAMDA´s model forecasts	
Real Behaviour		Online purchaser	Pure Off line Purchaser
	Online Purchaser	SUCCESS	ERROR
	Pure off line Purchaser	ERROR	SUCCESS

As described in the previous experiment, LAMDA has the capability to either assign each customer in just one segment or to assign each customer to each segment according to the GAD to each segment. Based on that, the two possibilities are considered for the validation stage.

Maximum GAD is considered (non overlapping)

The first type of validation did not consider the possibility of overlapping between segments (See Table 20).

Table 20 Measuring LAMDA´s forecasts (no overlapping)

Segment	**Number of customers located by LAMDA in each segment**	**Distance buying clients located in the segment**	**Number of Real Internet buyers within the segment (from V29)**	
1’	1640	67	76	**73.78%** (76/103)
2’	357	8	19	**18.44%** (19/103)
3’	66	3	8	**7.76%** (8/103)
TOTAL	2063	78	103	100%

Segment 1’ was proposed by LAMDA to be the one which joined the future online customers. Results show that the majority of the real online customers (73.78%) were located in this segment.

Forecasting accuracy was validated. As shown (See Table 21) LAMDA success rate was 73.78%. LAMDA identified 76 potential candidates among the 103 real online customers. However LAMDA also failed to recognize 27 real online customers as potential candidates.

Table 21 Interpretation of numerical results (no overlapping is considered)

	LAMDA's online purchasers	**LAMDA's pure offline purchasers**
Real online buyer	73.78%	26.20%

However, results in Table 21 were based on the maximum GAD in one segment but not with the possibility to locate the same customer in more than one segment.

Multiple GADS were considered (overlapping)

The capability of the model to assign each customer to each pattern of behaviour with its own GAD was taken into account for this second validation. In particular, for testing the results based on this overlapping of behaviours, customers with similar adequacy degree in more than one segment were assumed to belong to both segments simultaneously. As mentioned in the previous experiment, a similar adequacy degree is considered when the difference between the GAD of the same customer in each segment is inferior than 0.015. Taking this simultaneity into consideration, a new matrix resulted.

Table 22 Measuring LAMDA´s forecasts (overlapping)

Segment	**Number of customers located by LAMDA in each segment**	**Distance buying clients located in the segment**	**Number of Real Internet buyers within the segment (from V29)**	
1’	1640	67	90 (76+12+2)	**87.37%** (90/103)
2’	357	8	19	**18.44%** (19/103)
3’	66	3	8	**7.76%** (8/103)
TOTAL	2063	78	103	100%

Table 22 shows that 12 of the real internet customers located in segment 2’ presented a high GAD to segment 1’. Therefore, it can be interpreted that these 12 customers had the same possibility to behave as online customers as offline customers. They were assumed to possibly behave differently, according to the situation. Just two of the real internet buyers located in segment 3’ behaved in the same way that the customers in segment 1’. Based on that, the numerical interpretations of results when overlapping was considered are summarized in Table 23.

Table 23 Interpretation of numerical results (overlapping is considered)

	LAMDA's online purchasers	LAMDA's pure offline purchasers
Real online buyer	87.37%	26.20%

The number of customers who had bought online at SUPSA was still low (103), but at that time, official Lleida's Internet figures not only showed a 23% of Internet access between the citizens of Lleida but also a 10% purchasing rate. However, based on the real number, it is demonstrated that LAMDA forecasting accuracy increases when the multiple GADs are considered.

Conclusions

This experiment is the first research study that has used unsupervised learning techniques to forecast online purchasing based on in-store data. It has demonstrated the capability to forecast customer's (household) future behaviour from the secondary data collected from internal company's databases. Supported by judgmental forecasting and LAMDA's quantitative forecasting method, the implementation has been carried out. Moreover, the forecasting success has been assessed by comparing the forecast with the reality.
The most crucial stage when implementing the unsupervised forecasting model is the experts' participation. It is evident that the LAMDA´s unsupervised learning approach is more human-expert dependent than the supervised approach.

The experiment demonstrates that a predictive variable is required. There is no way to track the classification as LAMDA is a black box. Consequently, it is essential to count on a relevant variable that, although it is not exactly the same predictor, it has a high relationship with it.

Furthermore, the possibility to forecast online purchasers also identified whether the online customers are going to continue buying off line (the overlapping cases) is demonstrated in the experiment as not only the extreme behaviours are identified but also the ambiguous ones. From a managerial perspective, this experiment has introduced a new way to interpret the results to support decision making in marketing. Particularly, the fact that each customer presents a specific adequacy degree to each segment affects the traditional way of targeting.

Based on this, the interpretation stage is a key point. Each customer is assigned to one segment. However, there is also the possibility to assign each customer to each segment. This capability of the model has an advantage which consists of interpreting the results by a non-crisp point of view. All the customers who do not always behave in the same way may be identified. Therefore, the theoretical topic that the customer is likely to behave differently according to the situation is currently quantified in this experiment.

Despite the results being quite encouraging for future research, a high number of online customers would be needed to have a realistic measure of the forecasting success ratio of the LAMDA's unsupervised forecasting model.

References

AECE. "Resumen del Estudio sobre comercio electrónico B2C." Web page, 2003 [accessed December 2004]. Available at http://www.aece.org.

Aguado, J. C. "A Mixed Qualitative-Quantitative Self - Learning Classification Technique Applied to Simulation Assessment in Industrial Process Control." Universitat Politecnica de Catalunya, 1998.

Aguilar-Martin, J. and López de Mántaras R. "The Process of Classification and Learning: the Meaning of Linguistic Descriptors of Concepts." *Approximate Reasoning in Decision Analysis* (1982): 165-175.

Aguilar-Martin, J. and Piera N. "Les Connectifs Mixtes: de Nouveaux Opérateurs d'Association des Variables dans la Classification Automatique avec Apprentissage." In *Data Analysis and Informatics,* E. Diday et al., Elsevier Science Publishers, 1986, 253-265.

Bellenger, D. N. Robertson D. H. and Greenberg B. A. "Shopping Centre Patronage Motives." *Journal of Retailing* 53, no. Summer (1997): 29-38.

Berrell, J. "Just-in-Time Shoppers." *Retail World* 48, no. 12 (1995): 26.

Breitenbach, C. and Van Doren D. "Value-Added Marketing in the Digital Domain: Enhancing the Utility of the Internet." *Journal of Consumer Marketing* 15, no. 6 (1998): 558-575.

Bucklin, L. and Gautschi D. *The Importance of Travel Mode Factors in the Patronage of Retail Centres*. New York: North Holland, 1983.

Bucklin, R. E. and Sismeiro C. "A Model of Web Site Browsing Behaviour Estimated in Click stream Data." *Journal of Marketing Research* 40, no. 3 (2003): 249-267.

Burke, R. R. "Technology and the Customer Interface: What Consumers Want in the Physical and Virtual Store." *Journal of the Academy of Marketing Science* 30, no. 4 (2002): 411-432.

Casabayó, M. Agell, N. and Aguado, J.C. "Using AI techniques in the Groceryt Industry. Identifying the Customers most likely to defect." *The International Review of Retail, Distribution and Consumer Research* 14, no. 3 (2004): 295-308.

Chiang, W-y. K. Zhang D. and Zhou L. "Predicting and Explaining Patronage Behavior Toward Web and Traditional Stores Using Neural Networks: a Comparative Analysis With Logistic Regression." *Decision Support Systems* (2004): forthcoming.

Chiger, S. "Consumer Shopping Survey: Part 2." *Catalogue Age* 18, no. 11 (2001): 47-51.

Crawford, F. "Consumer Relevancy: Connecting With the 21st Century Market." *Global Online Retailing Report* (2000): 40-45.

Cymrot, A. Gelber A. and Cole L. L. "How to Invest in Office Building in Shopping Centers." *Real Estate Review* 12, no. Summer (1982): 75-82.

Dahlén, M. and Lange F. "Real Consumer in the Virtual Store." *Scandinavian Journal of Management* 18 (2002): 341-363.

Degeratu, A. M. Rangaswamy A. and Wu J. "Consumer Choice Behavior in Online and Traditional Supermarkets: The Effects of Brand Name, Price, and Other Search Attributes." *International Journal of Research in Marketing* 17 (2000): 55-78.

Douthu, N. and Garcia A. "The Internet Shopper." *Journal of Advertising Research* 39, no. 3 (1999): 52-58.

Downs, A. "A Theory of Consumer Efficiency." *Journal of Retailing* 37, no. 1 (1961): 6-12; 50-51.

Gehrt, K. C. and Yale L. J. "The Dimensionality of the Convenience Phenomenon: A Qualitative Re-Examination." *Journal of Business and Psychology* 8, no. 2 (1993): 163-180.

Howell, R. and Rogers J. "Research into Shopping Mall Choice Behaviour." *Advances in Consumer Research* (1980): 671-687.

Kalakota, R. and Whinston A. B. *Electronic Commerce, A Manager's Guide*. Massachusetts: Addison Wesley, 1997.

Kaufman-Scarborough, C. "A New Look at One Stop Shopping: a TIMES Model Approach to Matching Store Hours and Shopping Schedules." *Journal of Consumer Marketing* 13, no. 1 (1996): 4-25.

Kaufman-Scarborough, C. and Lindquist J. D. "E-Shopping in a Multiple Channel Environment." *Journal of Consumer Marketing* 19, no. 4 (2002): 333-350.

Kim, E. Y. and Kim Y-K. "Predicting Online Purchase Intentions for Clothing Products." *European Journal of Marketing* 38, no. 7 (2004): 883-897.

Mathwick, C. Malhotra N. K. and Rigdon E. "The Effect of Dynamic Retail Experiences on Experiential Perceptions of Value: an Internet and Catalogue Comparison." *Journal of Retailing* 78, no. 1 (2002): 51-60.

Montgomery, A. L. "Applying Quantitative Marketing Techniques to the Internet." *Interfaces* 31, no. 2 (2001): 91-108.

Oppewal, H. Louviewe J. and Timmermans H. "Modelling Hierarchical Conjoint Processes with Integrated Choice Experiments." *Journal of Marketing Research* 31 (1994): 92-105.

Padmanabham, B. Zheng Z. and Kimbrough S. O. "Personalization From Incomplete Data: What You Don't Know Can Hurt." *in Proceeding of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2001).

Raijas, A. "The Consumer Benefits and Problems in the Electronic Grocery Store." *Journal of Retailing and Consumer Services* 9 (2002): 107-113.

Ray, A. "How to Encourage Internet Shopping." *Marketing* May, no. 3 (2001): 41-42.

Reichheld, F. F and Sasser W. E. "Zero Defections: Quality Comes to Services." *Harvard Business Review* 68, no. 5 (1990): 105-111.Reimers, V. and Clulow V. "Shopping and Convenience: A Model for Retail Centres."*ANZMAC Conference*, 2000.

Schmittlein, D. C. and Peterson R. A. "Customer Base Analysis: An Industrial Purchase Process Application." *Marketing Science* 13, no. 1 (1994): 41-67.

Shim, S. and Drake M. F. "Consumer Intention to Utilize Electronic Shopping." *Journal of Direct Marketing* 4, no. 3 (1990): 22-33.

Shim, S. Eastlick M. A. and Lotz S. L. "Assessing the Impact of Internet Shopping on Store Shopping Among Mall Shoppers and Internet Users." *Journal of Shopping Centre Research* 7, no. 2 (2000): 7-43.

Strader, T. J. and Shaw M. J. "Characteristics of Electronic Markets." *Decision Support Systems* 21 (1997): 185-198.

Supphellen, M. and Nysveen H. "Drivers of Intention to Revisit the Web Sites of Well Known Companies." *International Journal of Market Research* 43, no. 3 (2001): 341-352.

Timmermans, H. Van der Heidjen R. and Westerveld H. "Cognition of Urban Retailing Structures: a Dutch Case Study." *Tijdschrift Voor Econ.En.Soc. Georgrafie* 73 (1982): 2-12.

Van den Poel, D. and Buckinx W. "Predicting Online-Purchasing Behaviour." *European Journal of Operational Research* 166, no. October (2005): 557-575.