Clustering Supermarkets: The Role of Experts

Armando B. Mendes[*]

Mathematical Department, Azores University,

R. da Mãe de Deus, 9501-801 Ponta Delgada, Portugal

Margarida G.M.S. Cardoso

Department of Quantitative Methods, Business School ISCTE,

Av. das Forças Armadas, 1649-026 Lisboa, Portugal

Abstract:

This work is part of a supermarket chain expansion study and is intended to cluster the existent outlets in order to support the evaluation of outlet performance and new outlet site location. To overcome the curse of dimensionality (a large number of attributes for a very small number of existing outlets) experts' knowledge is considered in the clustering process. Three alternative approaches are compared for this end, the experts being required to: 1- *a priori*: provide values for perceived dissimilarities between pairs of outlets; 2- *a posteriori*: evaluate results from alternative regression trees; 3- *interactively*: help to select base variables and evaluate results from alternative dendrograms. The later approach provided the best results according to the marketing experts.

Keywords: clustering; external validation; experts knowledge integration.

[*] Corresponding author

A supplementary exercise in cluster description involves the investigation of the clusters in order to establish whether or not they can be given substantive interpretations (…). Such substantive descriptions do not make direct use of data, but require investigators to reflect on the results of classification studies.

Gordon (1999)

1. Introduction

As in Europe, the retail sector in Portugal is going through a restructuring phase. Several authors (*e.g.* Birkin *et al.*, 2002, Dawson, 2000, and Seth and Geoffrey, 1999) identify such factors as increasing consumer mobility, increasing electronic commerce, changing household size, concentration of market power, home market saturation, and changes in planning legislation to justify the new trends in retailing. In the food retail, in particular, after an unprecedented period of hypermarkets growth, since the late 1970s, both in number and market share, it is now clear that hypermarket activity has slowed down significantly on behalf of the small or medium supermarkets (chain outlets including discount and hard discount chains) that nowadays present a larger dynamism.

In Portugal market share data shows that since 1996 the supermarkets were the only ones to grow simultaneously in the number of outlets and in the volume of sales and, consequently, to increase the market share from 28 to 34% in the A.C. Nielsen universe. In 1997 the supermarkets reached the leadership and consolidated its expansion strategy. According to the most recent data, in 2001, supermarkets' sales were already broadly superior to the sales in hypermarkets: 47% against just 35% of the total sales of outlets with alimentary products (see Figure 1).

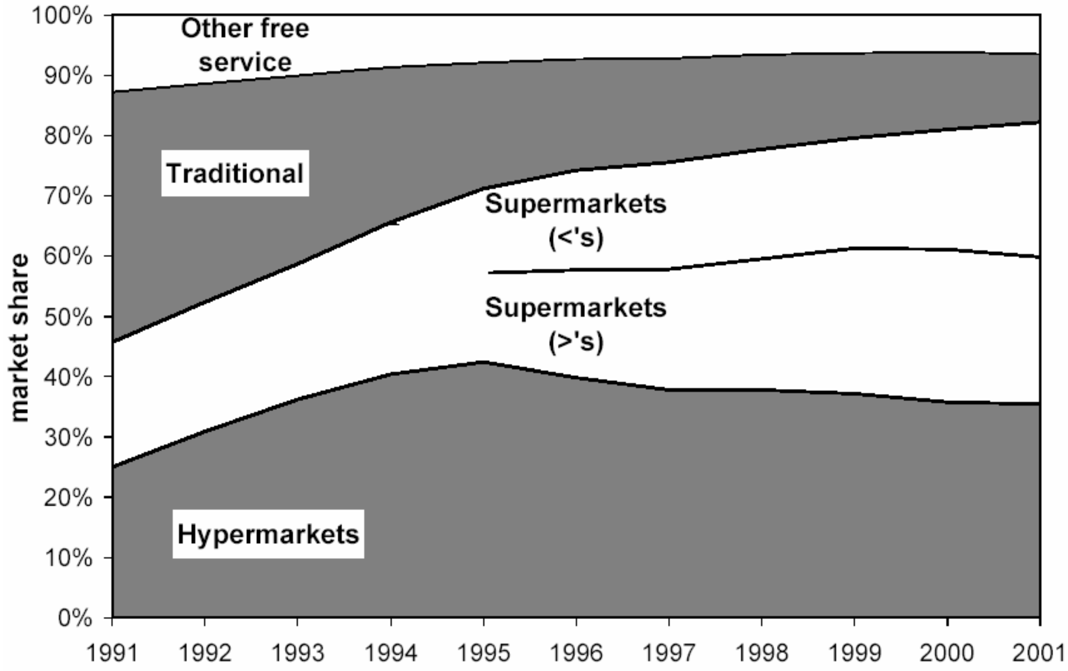


Figure 1

Market share evolution for food outlet type in the Portuguese market.

(Source: A.C. Nielsen Portugal)

This change in food outlet type is also found in other European countries (Birkin *et al.*, 2002). Much more stringent legislation and the fact that consumers are more demanding, force the retail groups to invest at outlets of smaller dimension, and so in a proximity and quality of goods and services strategy. This investment has a longer run return as well as smaller economies of scale, which forces careful decision-making (McGoldrick, 2000; Salvaneschi, 1996). Because smaller outlets are heterogeneous in aspects as location, dimension, and client behaviour, the definition of outlet clusters is essential in outlet performance and site evaluation.

2. Clustering supermarkets and the role of experts

This work is part of an expansion study of a supermarket chain with small and medium dimension outlets and is intended to cluster the existent outlets. The classification is not just useful to evaluate the relative performance of different locations and outlet management, but also to use in analogy forecast methods for the identification of potential site locations (Mendes and Themido, 2004). For that purpose several performance measures and other attributes were collected in a framework defined in this section. For addressing the high dimensionality of the data, the integration of expert knowledge in the clustering of supermarkets is suggested.

2.1. Measuring supermarkets' performance

The retailers soon realised the importance of outlet location, but understanding all the aspects of outlet performance, site locations, and the consumer's behaviour, forces to collect enormous amounts of information of several types as geographical, demographic, socioeconomic and regarding competition dynamics (Wedel and Kamakura, 2000; Themido *et al.,* 1998; Salvaneschi, 1996).

In order to organize all the data considered in location and outlet evaluation studies, an empirical classification of relevant variables is presented in Figure 2. This framework is intended for outlet and site evaluation of small to medium dimension outlets belonging to a retail chain, and is based in the authors' experience and in an extensive literature review. The variables are divided in three groups:

- The *location and outlet* attributes that are intended to evaluate aspects only dependent on the outlet and on the site location as the outlet characteristics, the accessibilities, and the image of the chain or the services range offered. Among the outlet characteristics, the commercial or sales area is the factor of major importance, which is emphasised in Figure 2 by an independent branch (Themido *et al.,* 1998, Salvaneschi, 1996).
- The outlets *influence area* attributes that are related with the evaluation of the trade area (or catchment area) generated by the outlet, which is essential in potential sales forecasting. These attributes are mainly demographic variables but also refer to the impact of the existent competition.

- *Clients' characteristics* that refer to their preferences, attitudes, behaviour, socioeconomic profile and geographic location are, finally, relevant in the evaluation of outlet performance.

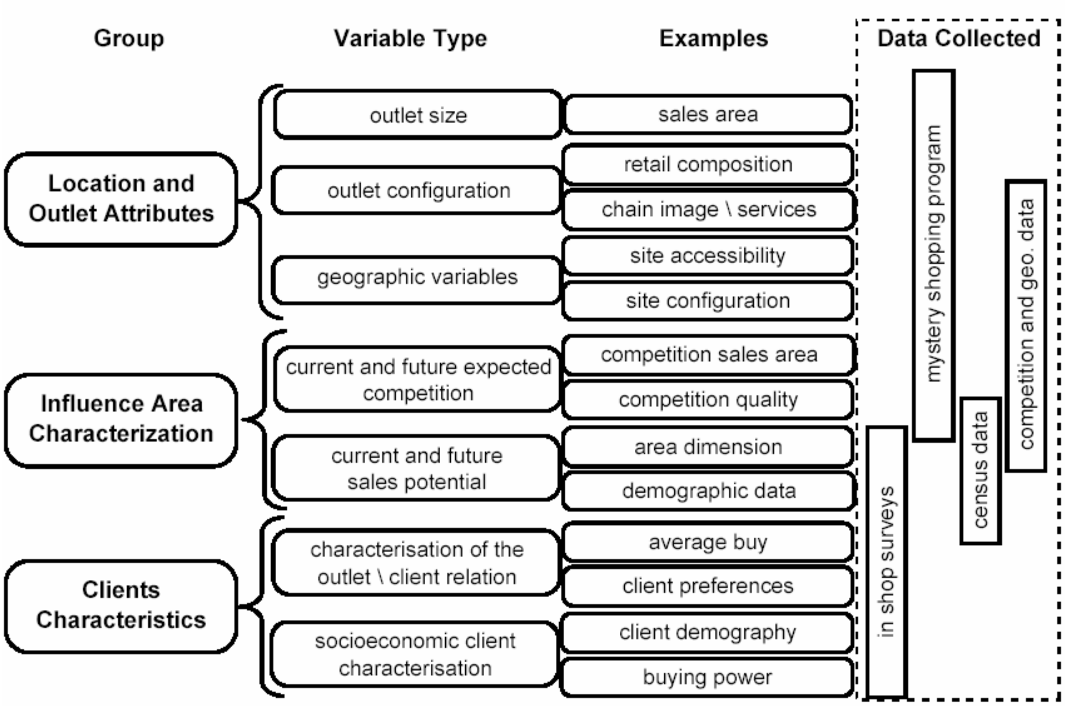


Figure 2

Classification of assessment location and outlet evaluation explanatory variables and data collected.

Few works attempted to classify outlet and site evaluation variables. One good exception is the work presented by Clarke *et al.* (2003), which is largely coherent with Figure 2. These authors used cognitive maps, based on answers of location experts from the largest retail chains in United Kingdom, to identify the main variables used in location decisions. This work confirms not only the suggested framework of the variables but also the high volume of data required in the outlet \ site evaluation studies.

2.2. The role of experts

In this work the clustering of outlets in a supermarket chain is performed. Since the clustering methods always impose a structure to the data, the validation of clustering structures (evaluation and comparison with other structures) is of major importance in order to accomplish the study objectives (Gordon, 1999).

Since quantitative internal validation reliability can not be assessed when only a small number of observations are available, the use of experts based external validation is essential (Wedel and Kamakura, 2000; Jain and Dubes, 1988, and Naert and Leeflang, 1978). In this work the use of expert knowledge is suggested for non quantitative external validation. The experts are marketing annalists' specialised in food retail outlet location, working with the supermarket chain since its origin and being responsible for all location and performance studies.

The use of expert knowledge, sometimes named domain knowledge, for evaluate the study quality is generally applied and investigated in areas as scale development (Hardestya and Bearden, 2004), marketing applications (Owrang, 2000, Pasa, 1996, Moutinho and Brownlie, 1994) and most relevantly in expert systems and automatic methods quality evaluation (Guijarro-Berdiñas and Alonso-Betanzos, 2002, Turban and Aronson, 2000, Adelman, 1991). Visual validation methods also imply some kind of expert or at least user assisted validation and interpretation (Hathaway and Bezdek, 2003, Hennig and Christlieb, 2002, Jones, 1996).

In the pattern recognition literature, Pedrycz (2004) mention the beneficial aspect of incorporating domain knowledge in the fuzzy clustering mechanisms. For justifying the use of this knowledge he suggests that a number of essential features may not be available or could not be easily quantified. Liu and Samal (2002) advocate that, by definition, the clusters represent same abstract concept that is clearly domain dependent.

In spite of that, few works have been presented considering the explicit integration of expert knowledge in feature selection and external validation for the clustering analysis (see Jain *et al.,* 1999, for a complete survey). One very good example is the utilization of a panel of experts to interpret classification rules, presented by Bay and Pazzani (2000). Their work concludes that many of the generated rules are useless or redundant and although it points for the subjectivity of the experts' interpretations, it confirms the need for this type of knowledge.

Several authors, as Liu and Samal (2002), and Halkidi *et al.* (2001), propose the use of external validation indexes for measuring the degree of agreement between expert delineated clusters and the ones obtained from a mathematical method. In this work the group of experts could not agreed in a cluster structure for the supermarket outlets, and considered that a difficult and subjective task. So, other approaches to expert knowledge integration are adopted, without asking for a cluster structure.

In the next sections the data collection phase is described and the three approaches are explained. In the *results section* these approaches are compared and a cluster profiling is presented. This paper finishes with conclusions and a methodological discussion.

3. Data Collection

A large number of variables were collected in order to account for the diversity of attributes that may influence outlets performance evaluation. This diversity of base clustering data is considered essential (see for example Wedel and Kamakura, 2000). Of all data collection procedures, explained in the next sections, a total of 250 variables were obtained, measured in all kind of scales, and covering all the aspects in the suggested variable framework (Figure 2).

3.1. In shop surveys

In shop surveys were hold in two different years during the study. The first took place in 2001 and was accomplished in all the existent supermarket outlets, in all days of two successive weeks, totalizing 3,766 valid questionnaires. The second was conducted in

2003, in a selected group of outlets and in selected days of the week, in a total of 2,394 valid questionnaires.

The questions included the clients' opinions regarding the configuration of the outlet, accessibilities and site configuration. They provided customers' demographic and socioeconomic characterization, attitudinal and behavioural attributes (motivation, means of transportation, choices and preferences) and the identification of the competition.

Quantitative variables as the *percentage of customers that come from home*, the *average monthly expenses in the outlet* or the *average value of purchase* were made available through the survey. When no significant differences were observed between the average values referring to 2001 and 2003, average values' yielding from the two surveys was used. In the rare cases where paired by outlet sample t statistic where lower than 5% (*e.g*. for the *mensal expenses in food*) the most recent value was used.

Client segmentation with data from both surveys was performed resulting in two segments. The segments were characterised and termed as *preferential costumers* and *eventual costumers* (Cardoso and Mendes, 2002). In consequence the *percentage of preferential costumers* was included as a new variable in the study.

3.2. The mystery shopping program

A **mystery shopping program** (*e.g.* Blankenship *et al.,* 1998), was accomplished with a visit to the outlet of an incognito analyst that observed visible aspects, did a buy, and evaluated several aspects in ordinal scales. These *in loco* observations were performed in all the chain outlets and in some of the most important competing shops.

They used a check list with several location attributes, outlet characteristics, accessibilities, outlet visibility, and some related with competition and characterization of the *influence area.* The variables are mainly nominal but some are subjective evaluations of some aspects of the outlet and service in an ordinal scale of nine points.

.

Coordinated with the mystery shopping program the outlets GPS coordinates were also collected along with their nearby competitors. This GPS coordinates and the mystery shopping data were loaded in a Geographical Information System and used to define influence areas, and to calculate variables used in the outlet characterisation and clustering.

3.3. Census, geographic and competition data

A large number of quantitative variables are available from the national geographical base of census 2001 data. This is high quality demographic data, accessible in several disaggregation degrees, and ready to use in a Geographical Information System. To include this data, influence areas must be defined along with criteria for geospatial intersection between these areas and the demographic areas.

Influence, trade or catchment areas can be defined as an area around the outlet from where it is likely to draw clients. Several methods have been suggested for its delimitation (*e.g.* Boots, 2002, Birkin *et al.*, 2002; McMullin, 2000), in the present case, shortest paths polygons and multiplicative weighted Voronoi diagrams were applied (Figure 3). The latter method allows, simultaneously, to incorporate the outlet attractivity and the competition in the outlet proximities (Boots and South, 1997). A data base with the location of more than 600 food outlets in Portugal was necessary for the method implementation.

Using space interaction procedures, percentage values and densities were calculated for all existing outlets. By the end of this process, and in spite of having made a careful selection of the census data, it resulted in more than half a thousand variables. To reduce this number the Pearson correlation coefficient matrix were calculated and the variables with more very strong correlations were deleted.

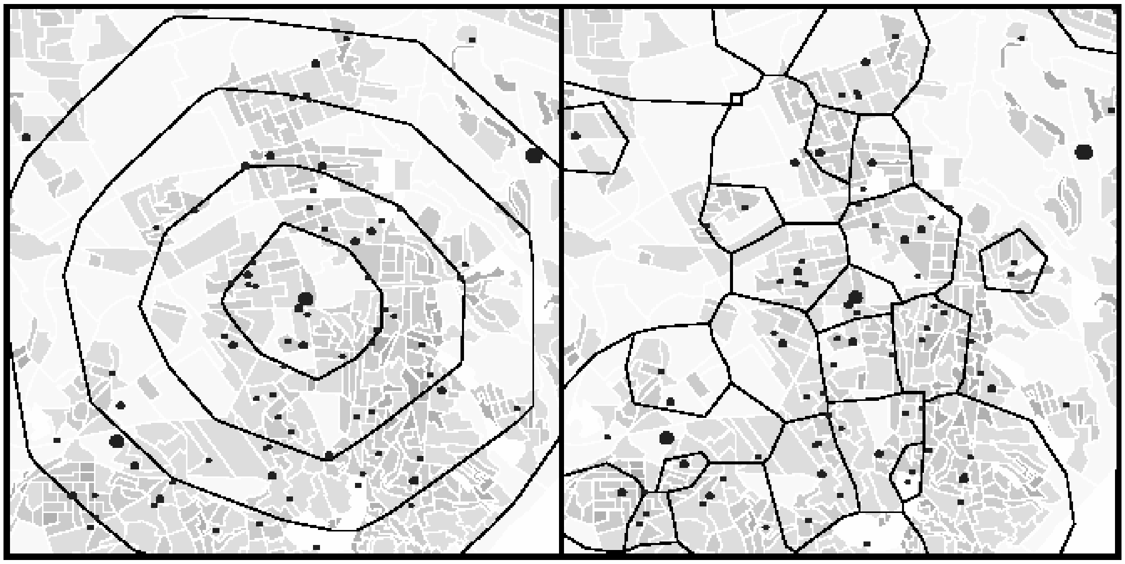


Figure 3

Shortest path polygons (left) and multiplicative weighted Voronoi diagrams (right) examples. (Point radius proportional to the outlet sales area and demographic polygons shaded by population density)

4. Methodological Approach

In spite of the data abundance following from last section, the number of outlet supermarkets was very small. This fact hindered the process of variable choice for the outlet clustering and respective characterization. To overcome this difficulty, three different procedures were considered for experts' knowledge integration:

1. In ***a priori*** integration approach the experts were required to compare pairs of outlets and evaluate their dissimilarities in a perceptual scale. The dissimilarities matrix was then directly used in the Ward's hierarchical clustering procedure. Finally, the selection of clusters' profiling variables relied on regressions over MDS dimensions.
2. In ***a posteriori*** approach the integration was accomplished by evaluating alternative clustering structures derived from a supervised learning procedure using regression trees. In this approach the base variables' choice relied on the regression tree procedure although experts required diverse regression tree parameterizations and target variables.

3. In the last procedure an interactive process was adopted: the experts' knowledge was considered in successive stages regarding the choice of clustering variables and the evaluation of clustering results. This process is termed ***interactive*** integration. Ward's Hierarchical procedure was based in alternative sets of base variables in order to provide "better" clustering, as judged by experts.

4.1. *A priori* experts' knowledge integration

In this approach the integration of the experts' opinion was made by means of outlet paired comparisons. The experts were requested to fill a questionnaire where pairs of outlets were compared and evaluated according to a scale of ordinal dissimilarities (from 1= *very similar outlets* to 9=*distinct outlets*).

The comparison was meant to be generic, although some aspects as location, management performance and site as well as clients' characterization were emphasised. The resulting symmetrical dissimilarity matrix was obtained by consensus among the several experts involved. This procedure is termed ***a priori*** as the experts opinion only regards the dissimilarities matrix used as clustering base. Clusters where then obtained using the hierarchical Ward's method resulting in the dendrogram presented in Figure 4. It should be noted that other hierarchical methods as centroid and group average linkage were tried and similar dendrograms were obtained.

From the observation of the fusion index values and fusion index variations *vs.* number of clusters chart, presented in the same Figure, six clusters were adopted. More complex stopping rules for determining the number of clusters were considered (see very good texts in Everitt *et al.,* 2001, and Gordon, 1999). Although conflicting results were obtained, in general, the six clusters cut was supported.

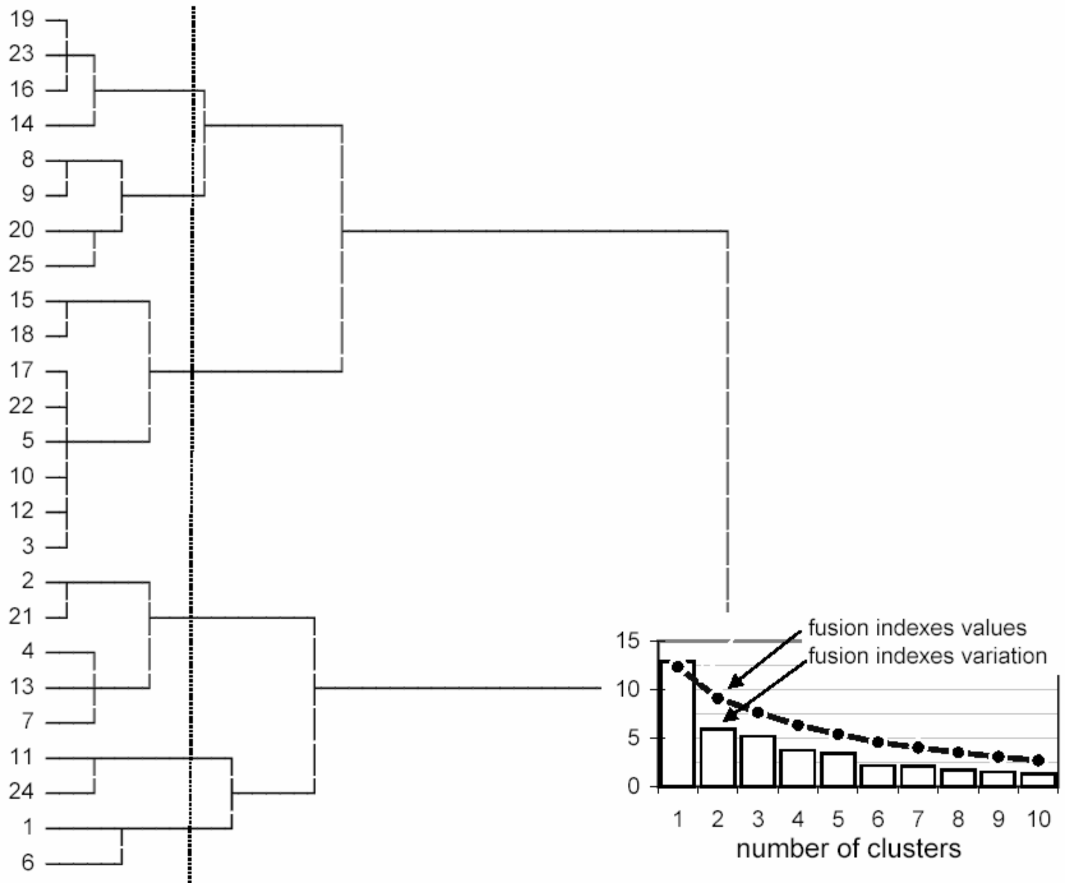


Figure 4

Dendrogram of the Ward hierarchical method directly applied to the dissimilarity matrix and fusion indexes chart.

In order to interpret the obtained clusters some further exploration of the dissimilarity data was required. MDS - Multidimensional Scaling non-metric analysis, using the ALSCAL algorithm by Takane, Young and Leeuw (Cox and Cox, 2000) was performed. A solution with four dimensions was found which accounts for an RSQ of 96% (Kruskal stress value of 7.8%). Figure 5 illustrates the positioning of chain outlets clusters in the extracted MDS dimensions, along with labels based in the clusters´ characterization.

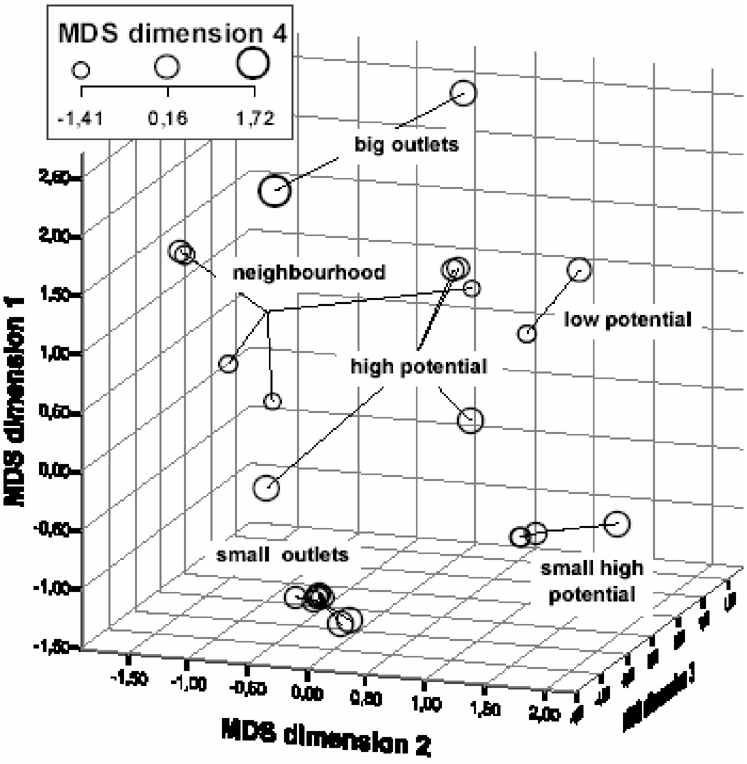


Figure 5

Outlets in the space of the MDS extracted dimensions and profile labels.

Regression procedures using several hundreds of target variables were performed considering the MDS dimensions as explanatory variables. Results enabled the identification of the variables responsible for the dissimilarity values among the outlets, depicted in Figure 6.

From this Figure *Dimension* 1 is related to outlet dimension and car parking facilities and inversely related with outlet visibility and sales per outlet area. *Dimension* 2 is related to influence area and percentage of households with children or working in primary or secondary sectors.

MDS *dimension* 3 is related with the number of elder residents and preferential costumers in the influence area, and *dimension 4* is associated with the percentage of occasional clients and complex trip (passage) clients. These results helped to support the clusters' profiling which may be summarized as follows (see Figure 5 and Figure 6):

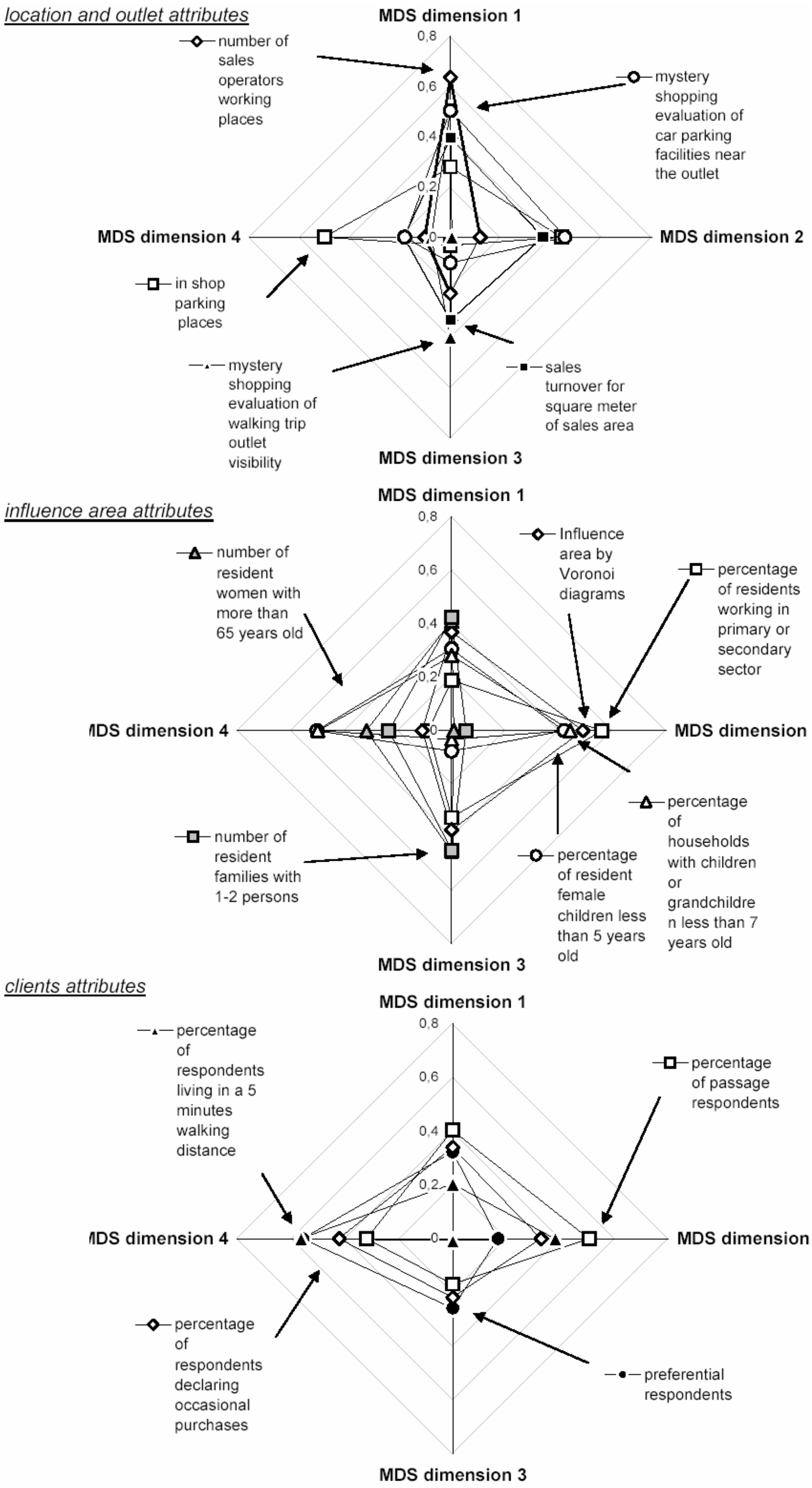


Figure 6

MDS dimensions characterization based on standardized absolute value regression coefficients (black marks represent attributes negatively related with MDS dimension).

- **Small outlets** a very homogenous cluster of eight outlets characterized by very negative values in *dimension* 1 and 3. According to the characterization of these dimensions these are small outlets with small influence areas indicating high levels of competition.
- **Neighbourhood outlets** constitute a five outlet cluster primarily characterized by low values in *dimension* 4 related to percentage of passage clients and low values in *dimension* 2 related to influence area dimension. So this cluster has the smallest influence areas and percentage of households with children and also very few occasional and passage clients.
- **High potential outlets**: four outlets with low to medium values in *dimension* 2 and high values in *dimension* 3. In consequence corresponds to outlets with small to medium influence areas but high percentages of preferential costumers justifying the high potential label.
- **Small high potential outlets**: a three-outlet cluster with very high values in MDS *dimension* 3 and very law values in *dimension* 1. This is a very high potential cluster with many preferential costumers and percentages of households with children but very small sales area.
- **Big outlets**: two very big stores as the high values of *dimension* 1 confirm. Both have negative values in *dimension* 2 indicating high levels of competition and small influence areas, and high values in *dimension* 4 indicating also high percentages of passage clients.
- **Low potential outlets:** two medium size outlets with high values in dimension 2, very law values in dimension 3 and low to medium values in 4. So these are outlets with big Voronoi influence areas, and consequently low levels of competition, and especially low levels of preferential costumers but also mean to low levels of passage clients.

4.2. *A posteriori* experts' knowledge integration

In the second approach experts' knowledge integration is made *a posteriori*, in the evaluation of alternative results provided by a regression tree method: CART **-** Classification and Regression Trees (Breiman *et al.,* 1984). Regression trees simultaneously cluster outlets and forecast the outlet turnover based on the target mean values in the tree leafs. Recent decision tree marketing applications can be found for instance in Cooley (2002), Blamires (2002), Micheaux and Gayet (2001), and Chou *et al.* (2000).

Alternative target variables were considered: the sales turnover for several years and the ratio of sales turnover over the sales area that is a very common outlet performance measure in the literature (see for example Birkin *et al.*, 2002). All the remaining available variables were considered as predictors. Several trees with different parameterisations were grown. . In the case of ties in variable selection for a splitting node, which were very common, both trees were grown and joined to the selection set.

Supervised learning methods rely on enormous amounts of data for internal validation (Berry and Linoff, 1997). In the present application the reduced number of outlets limited the use of the usual precision indexes when comparing alternative decision trees. However, indexes as leave-one-out estimates were presented to the marketing analysts for tree selection decision support.

In the tree selection and comparison process trees were rejected when counterintuitive decision rules emerged, for example if a bigger sales area corresponds to a leaf with mean smaller annual turnover

In Figure 7 the best tree is presented. This tree was evaluated by experts as *very good* since all the splits made sense and the clusters in the terminal nodes were also considered reasonable.

Clusters' profiling may be directly derived from the tree, which was the most appreciate characteristic of this approach as it greatly facilitates expert cluster validation. Thus, clusters were named according to the propositional rules associated with the corresponding leaf nodes (see Figure 7):

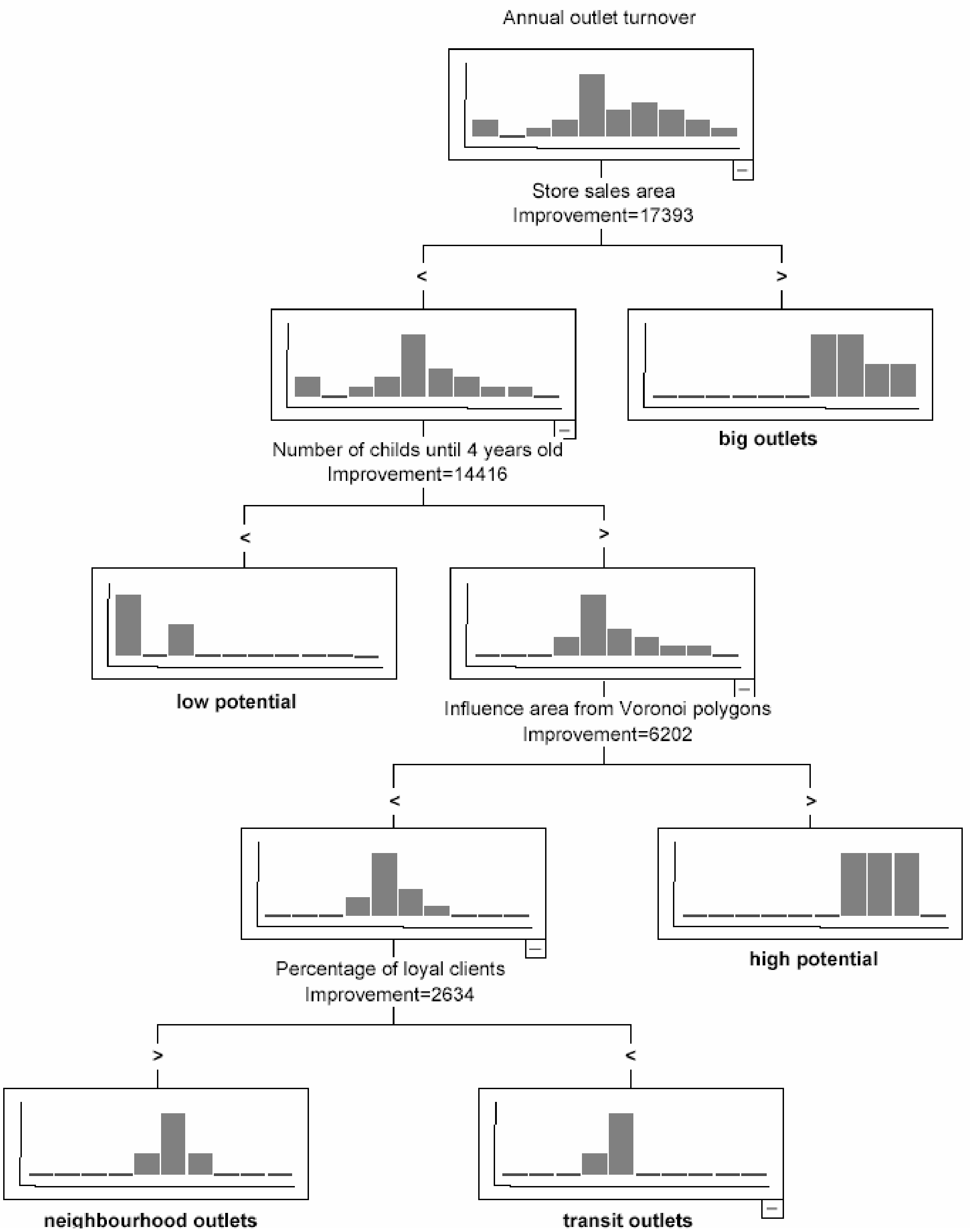


Figure 7

Selected tree obtained for the CART method.

(The bar graphs represent the histograms of the target variable in each node)

- **Big outlets** correspond to the bigger values for the outlets *sales area* and also to the biggest outlet *annual sales*.
- **Low potential** outlets have small to medium *dimension*, have small *number of children in the influence area* and very low *sales turnover*.
- **High potential** outlets are characterized by all the latter splits and *big influence areas by the Voronoi polygon methods* and correspond to *bigger sales*. It should be noted that the biggest outlets are excluded from this cluster as were split in the first tree node. And so, these are not necessarily de biggest influence areas for the existent outlets.

The last split variable was calculated as the percentage of inquire respondents which claimed to spend at least 75% of the mensal expenses in food on the outlet and the rest in a hypermarket. As we found that these loyal costumers had residences near the outlet, this cluster was called **neighbourhood outlets** and the other **transit outlets**.

4.3. Interactive experts' knowledge integration

In this approach the experts' knowledge was used in the base clustering variables selection as well in the appreciation of the results from the successive hierarchical clustering procedures. The process was reinitialised several times with new base clustering variables when the clusters didn't correspond to the expert's expectations. A constant dialog was maintained and all the analysis was in close agreement with the experts.

According to the same experts a measure of the dimension of the outlet area or the sales turnover should be considered as base variable. In addition, a measure of the customers' residential *versus* customers in transit proportion should also be taken into account, since these two clients' types were, *a priori*, perceived as different in terms of mean purchase.

The first factor could be translated by the annual sales turnover, the area of the outlet or a ratio between them. After several testing the annual sales turnover was selected, as it tends to contain the largest relative dispersion. The choice of the variable to translate the second factor assisted, also, to a similar criterion. In consequence a new variable was defined that represents the percentage of clients on *exclusive trips* to the outlet, *i.e.* the ones that came from home and return home after the purchase.

The final selected clustering results from the Ward method, but it was internally validated by constructing countless dendrograms with several combinations of methods and distance measures, with only hierarchical order variations. Finally, the results were externally validated by the experts that agreed to the clusters formed with only minor remarks.

In Figure 8 the obtained clusters are introduced including the cluster labels based on the characterization presented in section 5.3. In this Figure, the two outlets in the bottom of the chart were identified as outliers. Both had been previously picked up by retailing experts as these outlets had poor performances and dreadful locations.

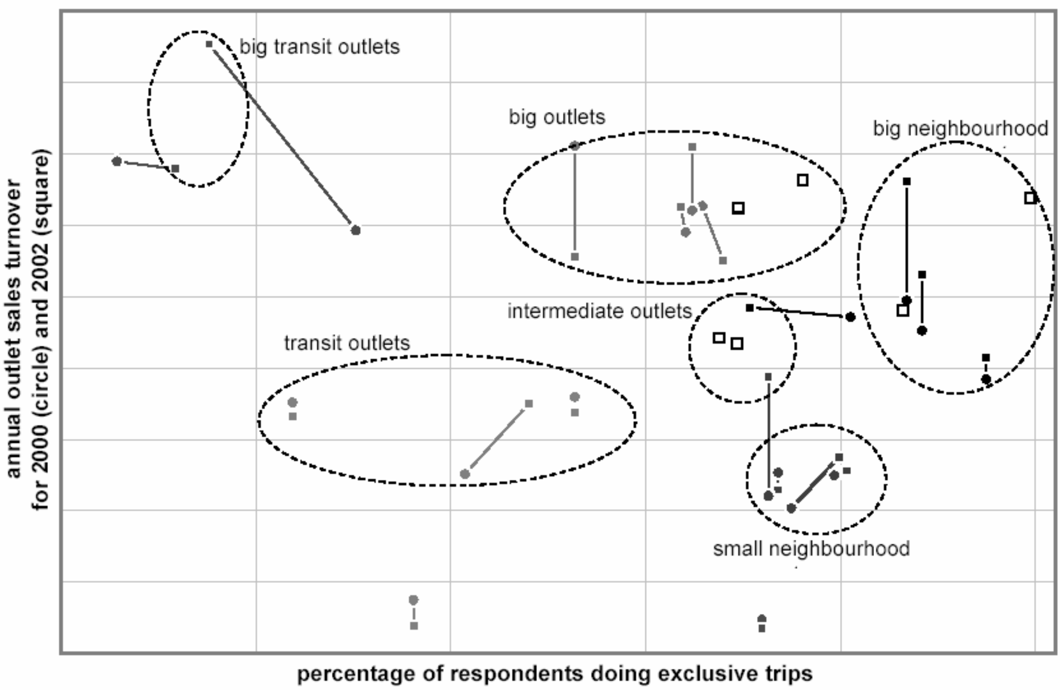


Figure 8

Clusters by interactive method showing two years of data.

(Empty squares represent new outlets in the two year period)

In the Figure 8 values refer to 2000 and 2002, as in these years in shop surveys were performed. In the latter year the inquiry was only done in some of the outlets, so a constant value were considered for plotting proposes. Empty squares represent six new outlets in 2002.

With the 2000 data only 4 clusters were identified. Between the two years two new clusters were formed. One of the new clusters is the **big transit outlets** which were formed by the convergence of two former outliers. The other one is the **intermediate outlets** being also formed by two outlets coming from other clusters and two new ones. This also emphasise the need for constant clustering revising as new outlets are open and new data are released.

5. Results Comparisons and Profiling

In order to reveal further differences between the cluster structures yielded from the three approaches, results were compared based in the *sales turnover dispersion* and in the *proportion of explained variance*. Finally, the supermarket clusters resulting from the interactive approach were profiled.

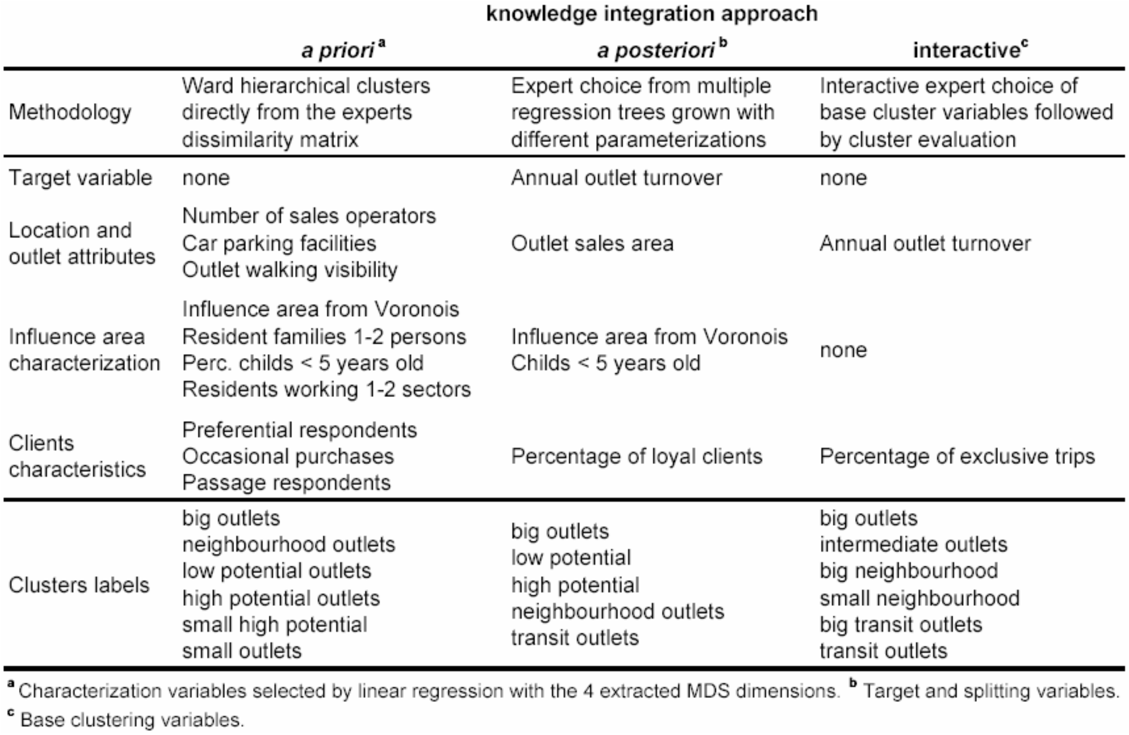
5.1. Sales Turnover Dispersion

In Table 1 the main characteristics of the different methodologies for expert's knowledge integration in the outlet clustering are summarised, clearly showing the diversity of approaches used. It should be noted that the *a posteriori* approach uses a supervised learning process while the others use unsupervised clustering procedures without any target variable. As it is shown, the base variables corresponding to the different methodologies are diverse.

In general the variables are well spread in the three empirical classification categories, suggested in section 2.1, meaning that the principal aspects empirically selected as necessary for outlet clustering and evaluation are supported in the results. One exception is the *interactive* approach where the *influence area* category hasn't any variable. This is a direct consequence of choice, by the experts, of only two variables as base cluster variables. In spite of that, the discriminant attributes selected by Chi-square and F tests, are very well spread for every variable category (see Table 3).

Table 1

Summary of the main characteristics for the 3 different methodologies.



	knowledge integration approach		
	a priori [a]	*a posteriori* [b]	interactive[c]
Methodology	Ward hierarchical clusters directly from the experts dissimilarity matrix	Expert choice from multiple regression trees grown with different parameterizations	Interactive expert choice of base cluster variables followed by cluster evaluation
Target variable	none	Annual outlet turnover	none
Location and outlet attributes	Number of sales operators Car parking facilities Outlet walking visibility	Outlet sales area	Annual outlet turnover
Influence area characterization	Influence area from Voronois Resident families 1-2 persons Perc. childs < 5 years old Residents working 1-2 sectors	Influence area from Voronois Childs < 5 years old	none
Clients characteristics	Preferential respondents Occasional purchases Passage respondents	Percentage of loyal clients	Percentage of exclusive trips
Clusters labels	big outlets neighbourhood outlets low potential outlets high potential outlets small high potential small outlets	big outlets low potential high potential neighbourhood outlets transit outlets	big outlets intermediate outlets big neighbourhood small neighbourhood big transit outlets transit outlets

[a] Characterization variables selected by linear regression with the 4 extracted MDS dimensions. [b] Target and splitting variables. [c] Base clustering variables.

The box plot charts for outlet annual sales turnover (Figure 9) help to illustrate further differences between the clustering results. In these charts the degree of cohesion of the different clustering results may be evaluated visually and outliers may be identified for the annual sales variables.

Although experts identified outliers were previously removed, Figure 9 reveals additional outliers (marked with circles and stars). In particular, the five outliers identified in the *a posteriori approach* can be justified by the tree parameterisation used. Since this parameterisation constraints a minimum of two outlets in each leaf, the presence of one isolated outlier could not be detected. Furthermore the very small data set can lead to outliers having high relative impact in impurity measures. However, no outlier could be identified in the box plots for the year used as target variable. This very good result is probably due to the retail annalist knowledge integration in the interpretation and evaluation of splits and clusters formed.

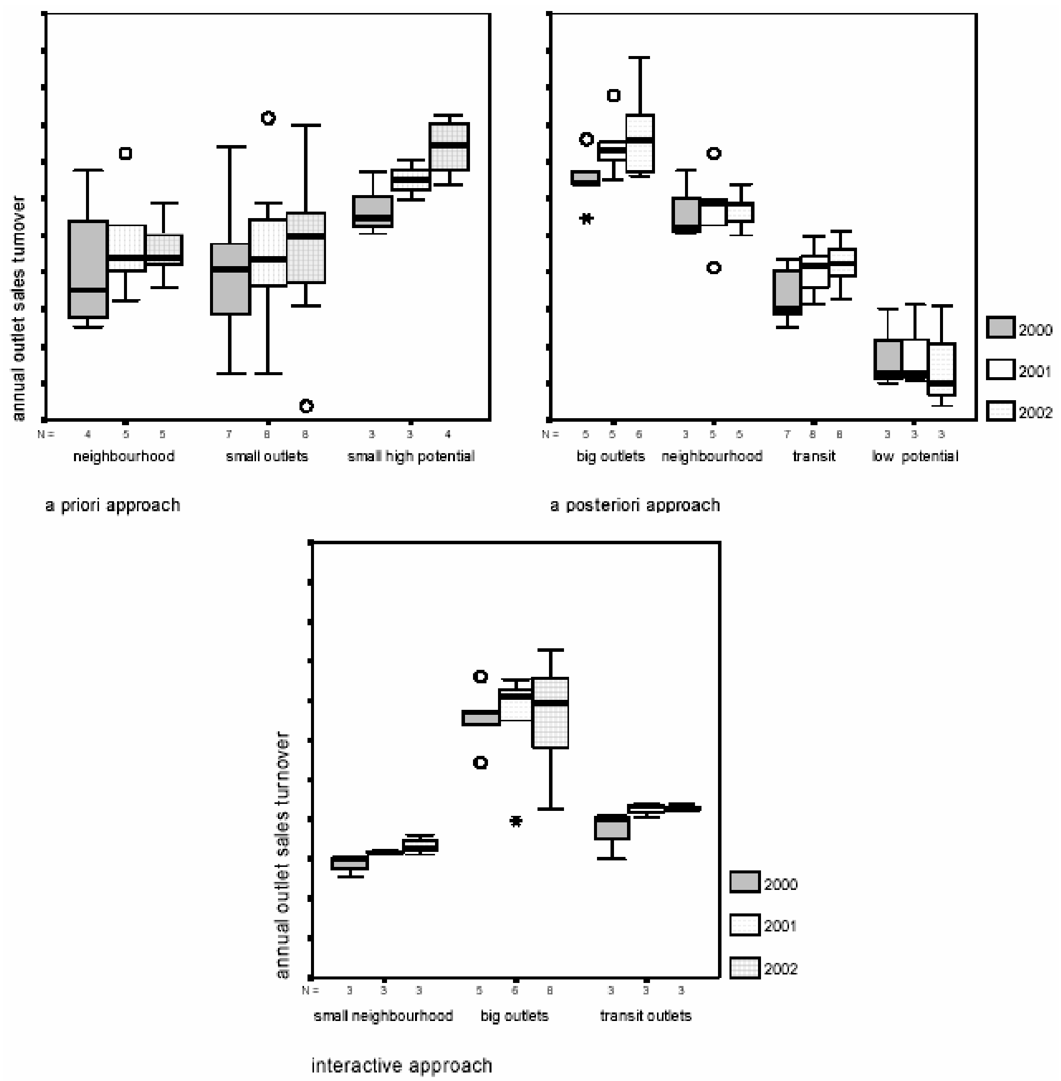


Figure 9

Box plot charts for annual sales individualizing clusters with three or more outlets. (Stars and circles are identified outliers, being circles 1.5 to 3 and stars more than 3 interquartile ranges).

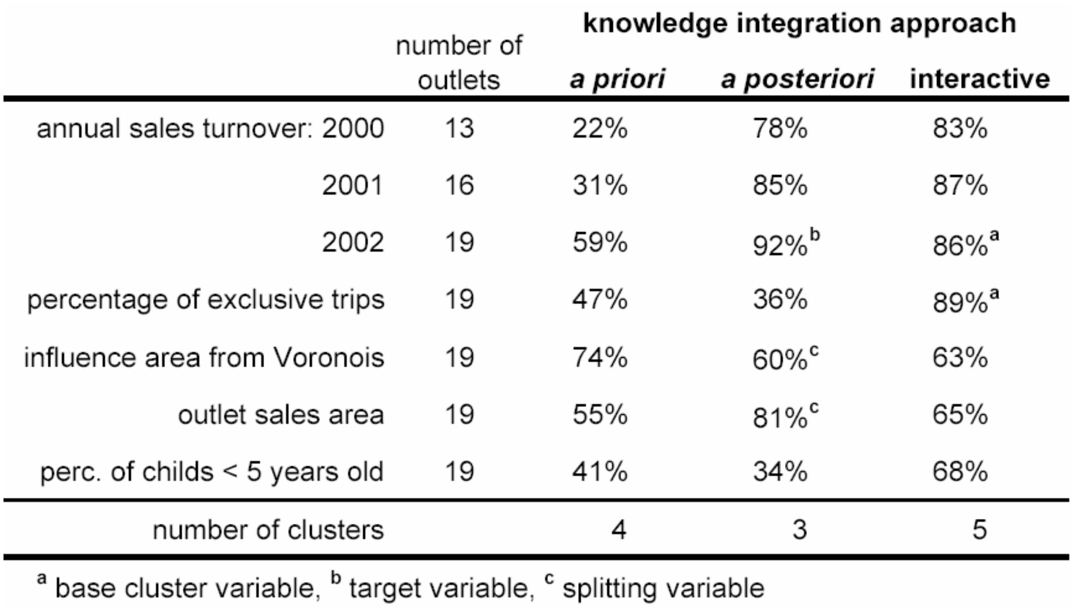
In what concerns dispersion of annual outlet sales the *a priori* approach presents the worst results. The better results refer to the *a posteriori* approach which is not surprising since the outlet sales 2002 turnover was used as target variable in a supervised learning procedure.

5.2. The Proportion of Explained Variance

In order to quantify the clustering *degree of cohesion,* the proportion of explained variance corresponding to the three approaches was calculated for some relevant variables (see Table 2). These variables were used either as base clustering variables (*interactive* approach) predictors (*a posteriori* approach) or simply variables that are strongly correlated with dissimilarities between outlets (*a priori* approach). For comparison purposes the intra-clusters variance was divided by the total variance calculated excluding the outliers. The same outlier outlets were considered for all approaches implying a cluster number reduction in two of them.

Table 2

Proportion of explained variance



	number of outlets	**knowledge integration approach**		
		a priori	***a posteriori***	**interactive**
annual sales turnover: 2000	13	22%	78%	83%
2001	16	31%	85%	87%
2002	19	59%	92%[b]	86%[a]
percentage of exclusive trips	19	47%	36%	89%[a]
influence area from Voronois	19	74%	60%[c]	63%
outlet sales area	19	55%	81%[c]	65%
perc. of childs < 5 years old	19	41%	34%	68%
number of clusters		4	3	5

[a] base cluster variable, [b] target variable, [c] splitting variable

From Table 2 the *a posteriori* and *interactive* approaches have similar results being the first better for the target variable but worst in other year's annual sales turnover. As the study objectives were to cluster outlets in homogenous but also time resistant clusters, it is not surprising that the results from de last approach were chosen.

When used in the lower part of the tree the splitting variables may not produce very high explained variance ratios. As they refer to only a limited number of outlets, the ones that were not discriminated in the preceding nodes, when the explained variance ratio for all outlets were calculated the values could be low, as is the case for the *influence area from the Voronois*.

In the retail location experts' opinion the *a priori* approach of knowledge integration was considered the least practical one as the high number of paired comparisons were considered "difficult". On the other hand, the location experts referred also the complexity of location and outlet evaluation as it involved a myriad of different evaluation aspects, and the difficulty in comparing outlets without the observation of quantitative data. This was probably the reason for the poor results observed.

The *a posteriori* and *interactive* approaches were much better evaluated by experts as they were not so demanding: they involved the comparison of different clustering results and choice of either the base clustering variables or the splitting variables for tree construction. Both approaches were considered easy to deal with and the *interactive* approach, in spite of being more time consuming, was considered very instructive and "actually a process of knowledge creation". The limitation of utilizing only one target variable was mentioned as the principal drawback of the *a posteriori* approach and the most valuated aspect was the easy to interpret trees and the ability to use variables in any scale quantitative or not.

5.3. Profiling Supermarket Clusters

In order to profile the supermarket clusters resulting from the *interactive* approach, Chi-square tests and F tests were used to support the existence of significant differences between the clusters for nominal and quantitative attributes, respectively. Only variables showing significant discriminatory power are considered in the following analysis.

In Table 3 a general view of clusters' characterization is presented which takes into account the variable framework presented in section 2.1. The quantitative variables were standardized by z-scores and for non quantitative variables relative frequencies were used.

Finally the main characteristics corresponding to each cluster are summarized considering the attributes and variables more relevant in each cluster. The attributes around mean values are usually not mentioned, but fundamental groups as performance indicators are always mentioned.

- **Big neighbourhood:** these are successful outlets as they assure the larger volume of sales per unit of outlet area. They are not located in downtown but in suburban zones of high potential and many residential households. The customers inhabit in close locations and frequently make exclusive trips to the store. They have above mean scholar qualifications and 73% were classified as preferential. The competition comes mainly from discount outlets and other chains.
- **Small neighbourhood:** these are the smaller outlets and also the ones with lower sales values. For the outlet configuration they present medium to reduced evaluations in almost all the parameters, and so a restyle is recommended. Car parking near the outlet is difficult and almost all clients came by foot. Their costumers are manly senior, spent high percentages of their budget in food in the outlet, and are almost all preferential clients. The competition cames mainly from discount outlets and small stores.

Table 3
Outlet cluster resume characterization.
(Arrows represent the most distinct vales for mean (vertical) and variance (horizontal) in each cluster, n\d is short form for not enough data).

	big neighbourhood		small neighbourhood		intermediate outlets		big outlets		transit outlets		big transit outlets	
location and outlet attributes - outlet performance and area												
2002 annual sales turnover			↓					⇔			↑	⇔
sales / area ratio	↑	⇔						⇔				
sales variation between 2002-2000					n\d	n\d		⇔	n\d	n\d	↑	⇔
sales area in square meters	↓		↓					⇔			↑	⇔
location and outlet attributes - outlet configuration												
cash machine in the outlet	↑		↑				↓		↓		↑	
price image from mystery shop.			↓			⇔			↑	⇔		
attendance sympathy			↓						↑	⇔		⇔
outlet tide \ cleanliness			↓	⇔	↑			⇔				
location and outlet attributes - geographic variables												
nº of outlets in downtown	↓				↑	⇔			↑	⇔	↑	
is the outlet considered anchor?							↑		↑		↑	
outlet walking visibility		⇔					↓	⇔	↑	⇔	↑	
parking facilities near outlet		⇔	↓		↓			⇔	↓			⇔
influence area characterization - competition												
also clients of other chains					↓	⇔		⇔	↑		↑	⇔
sum of competition sales area		⇔							↑	⇔		
outlet price relative to competition	↓	⇔						⇔	↑			
% of mensal expenses in outlet			↑			⇔			↓		↓	⇔
influence area characterization - sales potential												
number of households in influ. area	↑	⇔			↑	⇔				⇔		
nº non-residential builds				⇔	↑	⇔		⇔	↑	⇔		
nº of new buildings in last 4 years							↑	⇔				
clients characteristics - outlet \ client relation												
% exclusive trips to the outlet	↑								↓		↓	
% of clients walking to the outlet		⇔	↑			⇔					↓	
% of preferential costumers	↑		↑					⇔	↓	⇔	↓	
% clients living <= 5 min.	↑			⇔				⇔	↓	⇔	↓	⇔
clients characteristics – socioeconomic characterization												
% clients >= 45 years old		⇔	↑								↓	
% clients with outcome >= 1.600 €		⇔		⇔		⇔		⇔			↑	⇔
% clients <= primary school	↓										↓	

- **Intermediate outlets:** these outlets show medium values in all performance variables. They are located in smaller suburban towns, and have mean to high values in outlet configuration evaluation. They usually have easy accesses in walking trips but high parking difficulties. Although showing high variability, the influence areas present a fair potential with large number of households and non residential buildings. They also have a balanced equilibrium between preferential and eventual costumers. They suffer little competition from other similar chains and hypermarkets.

- **Big outlets:** this is the largest outlet typology and the most heterogeneous with high variability's in all performance variables. The number of outlets with cash machines is reduced since this group includes some of the oldest stores. Some have own parking places but the majority don't. These stores are often considered anchors of customers' attraction for the shopping centre or street. The influence areas present large dynamism since the number of buildings built in the last years is high. Clients came from both segments. The competition is generally high but variable from outlet to outlet.

- **Transit outlets:** these outlets have medium to low performance. But, they got good classifications in the outlet configuration and service. They are located in small shopping centres in downtown where they are considered attraction anchors. In spite of that, the parking facilities are poor, and the clients came from far away but, rarely, in exclusive trips. The influence area show high values of non residential buildings indicating working zones. These outlets are characterized by the eventual costumer segment, younger customers with higher incomes, and massive competition levels reflected in every variable.

- **Big transit outlets:** this small group had a very good performance in terms of sales turnover and a high growing tendency. They are located in big city centres where they are considered attraction anchors. The costumers spent only a reduced percentage of food expenses in these outlets. They move manly by car, came from distant places and rarely take exclusive trips. This cluster is also characterized by the eventual costumer segment, with younger costumers, and higher scholar qualifications. Competition levels are high coming from similar outlet chains and hypermarkets.

6. Discussion and Conclusions

When a large number of variables are available for clustering a small amount of observations, the need to integrate experts' knowledge in the clustering process becomes particularly relevant. In order to cluster a small number of supermarkets with a large number of available attributes three alternative approaches are presented which integrate experts' knowledge: the *a priori*, the *a posteriori* and the *interactive*.

According to the analysts' expectations the *a priori* approach should integrate the relevant experts' knowledge concerning the clustering of supermarkets as it is based on the perceived dissimilarities between the supermarkets. In the *a posteriori* approach experts' knowledge was required in order to select among alternative regression trees. Finally the integration of experts' knowledge both in the choice of base variables for clustering and in the selection of results was expected to give a larger role to the experts.

According to the experts' perspective some advantages and disadvantages of the three alternative approaches may be pointed:

- In the *a priori* approach the paired comparisons task was found to be very demanding and the results were poor.
- Regression trees used as a clustering tool in the *a posteriori* approach where found to be very attractive. Regression trees promoted the communication between the experts and the analysts as they simultaneously provide clusters and comprehensible descriptions.

- The interactive approach made the clustering process more transparent, leading to the chosen clustering results. It also allowed the identification of outliers. However, the process was considered to be very costly.

In the ***a priori* approach** sales related variables where expected to explain the perceived dissimilarities between the supermarkets since sales turnover is generally accepted as a major evaluation measure for comparing outlets performance. As it was not the case, some hypothesis may be raised which refer to the complexity of the comparative outlet evaluation task. In fact, as it was already stated, location and supermarket performance evaluation involves large numbers of attributes which may turn measures of perceived dissimilarities between supermarkets insufficient for clustering purposes. In order to better integrate diversity contained in the concept of *supermarket performance* several clustering base variables should be considered for selection, *the interactive* approach being more appropriate for this purpose.

From the ***a posteriori* approach**, experts where quite enthusiastic about the use of regression trees but, they did not pick its results to be the "best". In fact, this is a very instable approach when it refers to small data sets, which call for extremely careful external validation (Bay and Pazzani, 2000). However, considering that this clustering process was widely accepted by users, it should be further researched taking into account two main guidelines:

- The role of experts should be reinforced and should allow for interactive choice of surrogate variables.
- The choice of the appropriate target variable should be carefully conducted. For this end multiple criteria decision analysis may be considered in order to build a performance measure more adapted to expert's outlet evaluation. Alternatively, several trees with different target variables may be grown and the corresponding results combined in a consensus tree (see Lapointe and Cucumel, 2002, and Leclerc, 1998).

The **interactive approach** yielded the most satisfactory outlets' typology. Although being very time consuming this approach simultaneously invested in a trust construction process. Thus, the analysts concluded that results were easily accepted, as the experts understood the techniques strengths and weaknesses better. This approach minimizes what is known in Decision Support Systems terminology as the "black box effect" (Adelman, 1991) being similar to an expert visual validation methodology as the three-step-method mentioned in Hennig and Christlieb (2002), but tailored for a very high dimensional data set. Also Wang (2001) uses a similar procedure and identifies two supporting arguments. First it uses the entire data set, in contrast to cross validation methodologies, so that information is not lost. Second a satisfactory result can always be obtained in contrast to dead end procedures that offer non alternative result if the validation fails.

Several clustering base variables were considered for selection in the *interactive* approach, but only two variables were selected as base cluster variables. Although this may appear to be in conflict with the richness of information that could be considered to characterize the supermarkets, some remarks may be added:

- The two chosen variables are very different in nature being collected by distinct processes. They are also not correlated or related in any way.
- Several trials were made considering a larger number of base cluster variables but the experts could not find any improvement in the results.
- It can be argued that the use of more than necessary variables can be misleading as it can mask the existence of clusters in the data, introducing noise in the results. In fact, several authors (see Gnanadesikan, 2001, Milligan, 1996 or Gordon, 1999) underline the role of feature selection and extraction for clustering and argue that the bias should not be to include variables without additional information (Duda et *al.,* 2001).
- Additionally the remaining available data must be used in cluster interpretation and validation which is an absolutely necessary phase to confirm the

correctness of the defined typology and to characterize the groups (which should not be made with only base cluster variables).

In the present application, the small number of observations and the “curse of dimensionality” increased the relevance of experts’ knowledge integration in the process of clustering. According to this study experts’ knowledge integration should be considered in all stages of the clustering process, mainly in selection of base variables and also in the selection among alternative clustering results.

The supermarket typology that was obtained as a result is already being used for differentiating marketing actions. Thus the frequent gap between theory and practice was overcome and the last stage of the clustering validation process was reinforced.

References

Adelman, L., 1991. Evaluating Decision Support and Expert Systems. Wiley-Interscience, Chischester, UK.

Bay, S.D., Pazzani, M.J., 2000. Discovering and describing category differences: What makes a discovered difference insightful? In: Gleitman, L.R., Joshi, A.K. (Eds.), Proceedings of the 22th Annual Meeting of the Cognitive Science Society. Institute for Research in Cognitive Science, USA, pp. 603-609.

Berry, M.J.A., Linoff, G., 1997. Data Mining Techniques: For marketing, sales, and customer support. John Wiley & Sons, USA.

Birkin, M., Clarke, G., Clarke, M., 2002. Retail Geography and Intelligent Network Planning. John Wiley & Sons, Chischester, U.K..

Blamires, C., 2002. Segmentation. In: Birn, R.J. (Ed.), The International Handbook of Market Research Techniques. Kogan Page, London, U.K., pp. 497-518.

Blankenship, A.B., Breen, G.E., Dutka, A.F., 1998. State of the Art Marketing Research. NTC Business Books, Chicago.

Boots, B., 2002. Using local statistics for boundary characterization. In: Boots, B., Okabe, A. e Thomas, R. (Eds.), Modelling Geographical Systems: Statistical and computational applications. Kluwer Academic Publishers, Dordrecht, Netherlands, pp. 33-44.

Boots, B., South, R., 1997. Modeling retail trade areas using higher-order, multiplicatively weighted voronoi diagrams. Journal of Retailing 73 (3), 519-536.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Inc., California, USA.

Cardoso, M.G.M.S., Mendes, A.B., 2002. Segmenting clients from small supermarkets [Portuguese]. In: Carvalho, L., Brilhante, F., Rosado, F. (Eds.), Proceedings of the 9th Annual Congress of the Portuguese Statistical Society. SPE, Ponta Delgada, pp. 157-170.

Chou, P.B., Grossman, E., Gunopulos, D., Kamesam, P., 2000. Identifying prospective customers. In: Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining. ACM press, New York, USA, pp. 447-456.

Clarke, I., Mackaness, W., Ball, B., Horita, M., 2003. The devil is in the detail: Visualising analogical thought in retail location decision-making. Environment and Planning B 30 (1), 15-36.

Cooley, S., 2002. Loyalty strategy development using applied member-cohort segmentation. Journal of Communication Management 19 (7), 550-563.

Cox, T.F., Cox, M.A.A., 2000. Multidimensional Scaling. Chapman & Hall, U.K..

Dawson, J., 2000. Retailing at century end: Some challenges for management and research. The International Review of Retail, Distribution and Consumer Research 10 (1), 119-148.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. Wiley-Interscience, New York.

Everitt, B.S., Landau, S., Leese, M., 2001. Cluster Analysis. Edward Arnold Publishers, London, U.K..

Gnanadesikan, R., 2001. Cluster analysis: An overview of aims, aids, & challanges. In: Neves, M., Coelho, C.A., Cadima, J., Proceedings of the 8th Annual Congress of the Portuguese Statistical Society. SPE, Peniche, Portugal, pp. 39-57.

Gordon, A.D., 1999. Classification. CRC Press, Boca Raton, U.K..

Guijarro-Berdiñas, B., Alonso-Betanzos, A., 2002. Empirical evaluation of a hybrid intelligent monitoring system using different measures of effectiveness. Artificial Intelligence in Medicine 24 (1), 71-96.

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2001. On clustering validation techniques. Journal of Intelligent Information Systems 17 (2/3), 107-145.

Hardestya, D.M., Bearden, W.O., 2004. The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. Journal of Business Research 57 (2), 98-107.

Hathaway, R.J., Bezdek, J.C., 2003. Visual cluster validity for prototype generator clustering models. Pattern Recognition Letters 24 (9-10), 1563-1569.

Hennig, C., Christlieb, N., 2002. Validating visual clusters in large datasets: fixed point clusters of spectral features. Computational Statistics & Data Analysis 40 (4), 723-739.

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, USA.

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: A review. ACM Computing Surveys 31 (3), 264-323.

Jones, C.V., 1996. Visualization and Optimization. Kluwer Academic Publishers, The Netherlands.

Lapointe, F.-J., Cucumel, G., 2002. Multiple consensus trees. In: Jajuga, K., Sokolowski, A., Bock, H.-H. (Eds.), Classification, Clustering, and Data Analysis: Recent advances and applications. Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, Berlin, pp. 359-364.

Leclerc, B., 1998. Consensus of classifications: The case of trees. In: Rizzi, A., Vichi, M., Bock, H.-H. (Eds.), Advances in Data Science and Classification: Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98) Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, Berlin, pp. 81-90.

Liu, M., Samal, A., 2002. Cluster validation using legacy delineations. Image and Vision Computing 20 (7), 459-467.

McGoldrick, P., 2000. Retail Marketing. McGraw-Hill Europe, U.K..

McMullin, S.K., 2000. Where are your customers: Raster based modeling for customer prospecting. In: Proceedings of the Annual ESRI International User Conference. ESRI online Library, pp. 795-823.

Mendes, A.B., Themido, I.H., 2004. Multi outlet retail site location assessment: A state of the art. International Transactions in Operations Research 11 (1), 1-18.

Micheaux, A., Gayet, A., 2001. Turning a marketing database into a relationship marketing database. Interactive Marketing 2 (4), 327-346.

Milligan, G.W., 1996. Clustering validation: Results and implications for applied analyses. In: Arabie, P.; Hubert, L.J., De Soete, G., Clustering and Classification. World Scientific, Singapore, pp. 341-375.

Moutinho, L., Brownlie, D., 1994. The stratlogic approach to the analysis of competitive position. Marketing Intelligence and Planning 12 (4), 15-21.

Naert, P.A., Leeflang, P.S.H., 1978. Building Implementable Marketing Models. Kluwer Academic Publishers, Boston.

Owrang, M.M., 2000. Using domain knowledge to optimize the knowledge discovery process in databases. International Journal of Intelligent Systems 15 (1), 45-60.

Pasa, M., 1996. The value of marketing expertise. Management Science 42 (3), 370-388.

Pedrycz, W., 2004. Fuzzy clustering with a knowledge-based guidance. Pattern Recognition Letters 25 (4), 469-480.

Salvaneschi, L., 1996. Location, Location, Location: How to select the best site for your business. Psi Research - Oasis Press, Grants Pass, USA.

Seth, A., Randall, G., 1999. The Grocers: The rise and rise of the supermarket chains. Kogan Page, London, U.K..

Themido, I., Quintino, A., Leitão, J., 1998. Modelling the retail sales of gasoline in a Portuguese metropolitan area. International Transactions in Operations Research 5 (2), 89-102.

Turban, E., Aronson, J.E., 2000. Decision Support Systems and Intelligent Systems. Prentice Hall, USA.

Wang, S., 2001. Cluster analysis using a validated self-organizing method: Cases of problem identification. International Journal of Intelligent Systems in Accounting, Finance and Management 10, 127-138.

Wedel, M., Kamakura, W.A., 2000. Market Segmentaion: Conceptual and methodological foundations. Kluwer Academic Publishers, Massachusetts, USA.