SPATIAL ANALYSIS DATA IN SUPERMARKET SITE ASSESSMENT

Armando B. Mendes
Mathematical Department,
Azores University,
R. da Mãe de Deus, 9501-801
Ponta Delgada, Portugal
amendes@notes.uac.pt

Margarida G.M.S. Cardoso
Dep. of Quantitative Methods,
Business School ISCTE,
Av. das Forças Armadas,
1649-026 Lisboa, Portugal
margarida.cardoso@iscte.pt

Rui Carvalho Oliveira
CESUR, Inst. Sup. Técnico,
Lisbon Technical University
Av. Rovisco Pais, 1049-001
Lisboa, Portugal
roliv@ist.utl.pt

Abstract: In this presentation a methodology is suggested for site assessment and site selection based in a data analysis framework. This framework consists in tree steeps using different data analysis methods from cluster analysis, classification trees and regression analysis. The different variables selected in all the models used for the three steeps are compared and the spatial analysis data importance in site assessment is evaluated. For variable importance several measures can be used. In this text we use the discriminant power for the selection of profiling variables and a precision index for classification trees. Dominance analysis is used for the multiple regression models in the forecasting procedure. The advantage of this technique over other techniques as the standardized coefficients is that it overcomes some limitations of stepwise regression. The different measures of variable importance result in a clear pattern for the relevance of spatial analysis variables, being only dominated by the "trade area". So, spatial analysis resulting variables are of special importance in site assessment studies and the delimitation method used for their calculation is relevant in their usefulness.
Keywords: Spatial data analysis; Voronoi diagrams; Influence area; Discriminant analysis; Classification Tree; Precision Index; Jackknife validation; Dominance Analysis.
Communication Domain: Discriminant Analyses and Decision Trees
Application Domain: Economy, Management and Finance
Application Work in Progress

1. Data in Site Assessment

For site assessment and site selection decisions for chain supermarkets a suitable data framework is suggested where the data is classified in three groups namely: location and outlet attributes, influence area characterization and clients' characteristics. This empirical 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, and was already presented and analysed in Mendes and Cardoso (2005).
The theoretical importance of demographic data (census data) and other spatial analysis data (competitive and geographical data) is clear from that framework. Only the outlet size, outlet configuration and client characteristics are not covered by this type of data. But, for chain outlets the outlet configuration, and in some way the outlet trade area, tends to be very similar inside the chain, and the clients' characteristics can only be used in the cluster formation as they can not be used for forecasting proposes. In classification trees and regression analysis only mystery shopping and spatial analysis data can be used, and this data evaluates different aspects of store location assessment.
This data is used in a site assessment methodology by new store sales forecast models using data analysis methods as linear regression, cluster analysis and logical discriminants or classification trees. The method starts by the definition of analogue groups of existent outlets and their careful characterization. Subsequently, several alternative prepositional rules are induced by the identification of variables capable to discriminate among the different analogue groups. The objective is to classify a new potential site to one of the previous identified analogue outlet group. The last steep uses a linear regression model for new site sales forecast, which includes several predictors and dummy variables for the analogue group. In the Figure 1 a schema explains the entire process.

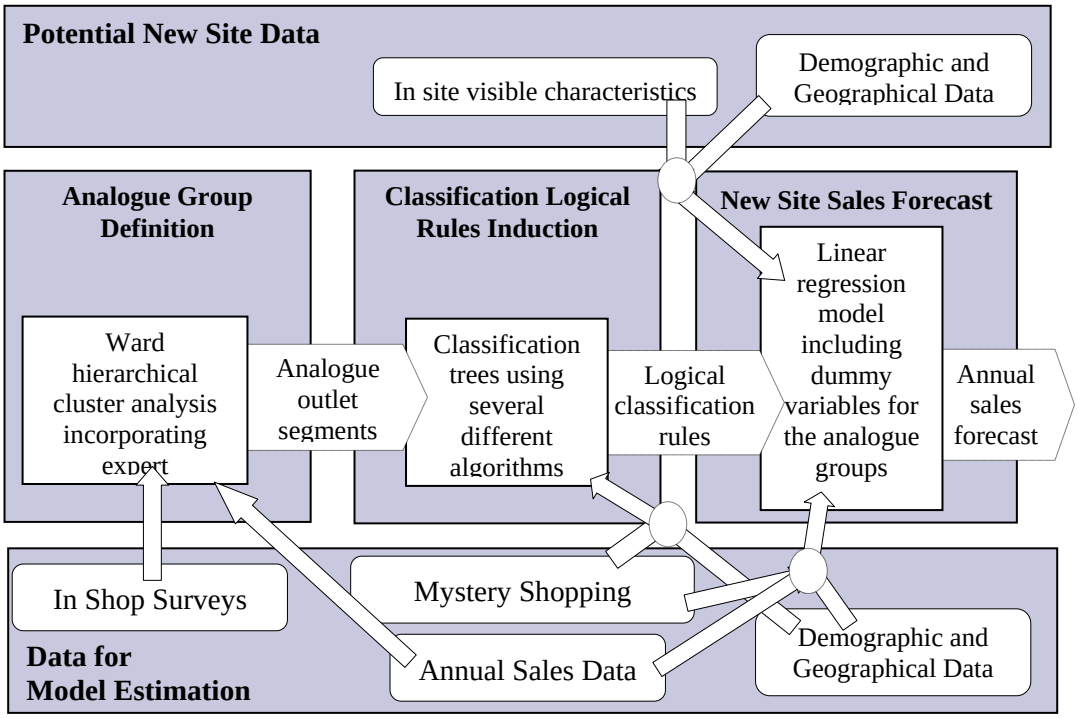

Figure 1 Schema explaining the data analysis analogue method for new site sales forecast and the data used for estimation and required for a potential site sales forecast.

A large number of variables were collected in order to account for the diversity of attributes that may influence outlets performance evaluation, and so, should be included in model estimation. This diversity of data is considered essential in data analysis methods. The data collection phase in the project was very time consuming and concerned several different techniques enumerated and explained in a previous work (Mendes and Cardoso, 2005). Of the fusion of all data collection procedures, a total of several hundred variables were obtained, measured in all kind of scales, and covering all the aspects in the variable framework for outlet location and site assessment.

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 in point studies as is the case of outlets, **influence areas** must be defined along with criteria for geospatial intersection between these areas and the demographic areas. Besides the calculation of demographic variables, influence areas are equally very useful for the definition of competitive pressure variables.

In the present case, Shortest Paths Algorithms (SPA) and multiplicative weighted Voronoi diagrams, first (MWVD) and second order (O2MWVD), were applied (Figure 2). The latter method allows, simultaneously, incorporating the outlet attractivity and the competition in the outlet proximities and is independent of any street network and frequently inaccurate average velocity data by street fragment (Boots and South, 1997). A data base with the location of more than 1300 grocery outlets in Portugal was necessary for the method implementation. This data was collected in coordination with the mystery shopping program and by recording GPS coordinates outside the outlets door. This GPS coordinates and the mystery shopping data were loaded in a Geographical Information System.

The calculation of demographic and competition variables also included the spatial intersection of those influence areas with the administrative limits of the statistical sections with associated demographic data. For the aggregation of the values of the several resulting polygons two different methods can be used. Authors as Cowen *et al.* (2000) and McMullin (2000) use the fraction of the statistical section covered by the influence area as weight in a weighted average. This procedure corresponds to the idea of uniform distribution of the data in the statistical section. Another available alternative consists of

using the same weight in an inclusion decision rule for the statistical section. In this work, the 50% limit value is used to include statistical sections with higher fractions area covered, and to exclude sections with lower fractions. This method has the disadvantage of distorting the original influence areas (compare shaded areas with influence area polygons in Figure 2), and the major advantage of adjusting the boundaries of the influence area to the boundaries of the statistical sections, which can be more appropriate as the statistical sections are defined considering geographical barriers.

Figure 2. Shortest path polygons (left) and multiplicative weighted Voronoi diagrams, first (centre) and second order (right), examples. (Stores as points and influence areas in grey. Road network in grey).

With the objective of comparing the different techniques used namely in the conjugation of the influence areas delimitation methods with the aggregation procedure, for the present case of chain grocery outlets, linear regressions were used using as explanatory all continuous variables resulting from the spatial analysis, calculated by the conjugation of the particular influence area delimitation method and aggregation procedure. Although the results presented in Mendes *et al.* (2004) refer to a small number of stores and cannot be widespread, they indicate a clear preference of the aggregation method for the decision rule over the weighted average. On the contrary, in relation to the delimitation model the preference is not clear, in this way the decision was to maintain the three models and to include in the study variables calculated by which one.

2. Dominance Analysis and the Importance of Spatial Analysis Data

In steep one of the methodology suggested (Figure 1) the discriminant power of the different variables were compared using the Kruskal-Wallis p-value and a nominal factor variable considering the six supermarket segments defined in Mendes and Cardoso (2005). For variables without the order feature Chi-square tests were used. F tests of variance analysis were not used for metric variables for two reasons. First, only a small sample of a couple of dozen outlets was available and so the Normality distribution and the constant variance assumptions are difficult to ascertain and, secondly the preference was endorsed to a common measure. Realizing the fact that ordinal variables can not be tested for discriminant power by a F test and the Kruskal-Wallis, being a non-parametric test, is considered very trustworthy the later was used whenever was possible.

All variable types and data collection methods emerge in the first third of the ranking list supporting the need of using all the different collection techniques and a plethora of

different variable types in outlet assessment and segmentation. In spite of this, the variables resulting from data analysis and classified as “influence area characterization - sales potential” are, in this case, clearly the largest group with nine variables very well classified in the ranking of discriminant power.

After the definition of groups of analogue existent stores and their careful characterization, the objective were to identify variables and propositional rules capable to discriminate among the different groups of stores. The leave-one-out was used as a tree validation methodology method as a particular case of jackknife validation or the U-method (Crask and Perreault, 1977), and consists of classifying each one of the outlets according to a tree built with the remaining ones. The error estimate corresponds to the number of erroneous classifications over the total number of outlets or trees built. This method estimates an error classification with some realism, when the number of observations is reduced (Lattin *et al.,* 2003, Gentle, 2002). The “best” models as evaluated by the experts using domain knowledge, present leave-one-out error estimates among 22% for the tree built from CART algorithm, 26% for the CHAID and 35% for QUEST. However, all the models are significant at 1% according to the test Q of Press, which evaluates the quality of the model using as null hypothesis an aleatory classification.

These values of error estimate for the whole logical model are used in the precision index along with two measures of propositional rule quality (Cardoso and Moutinho, 2003, Quinlan, 1993). As the leaf is attributed to the modal group and the number of outlets per group is very low, is desirable that only one leaf is attributed to any group, being the “percentage of outlets of the group in the leaf” a measure of the dispersion of the group for several leafs of the classification tree. On the other hand, the “hits percentage in the leaf” measures the degree of purity or the homogeneity of a leaf. Both are intended to maximization. Based in that assumptions, the expression (1) was introduced where the **precision index** for outlet *j* is represented by IP_j, *leaveOneOut* represents the estimate of the classification error by the leave-one-out method for the model (*a*), and the *%hits* the “hits percentage in the leaf” regarding the propositional rule and *%grupo* the “percentage of stores of the group in the leaf” for the same rule (a_r).

$$IP_j = \left(1 - leaveOneOut_a\right)^{\beta} \times \left(\%hits_{a_r}^{\alpha} \times \%grupo_{a_r}^{1-\alpha}\right),\ 0 \le \alpha \le 1,\ \beta \ge 1 \qquad (1)$$

Several variables are common between a ranking list by Kruskal-Wallis p-value and the precision index for classification rules, but many are different especially in the lower level of the trees. Both methods are based in discriminante methodologies, but in logical tree discrimination the use of variables in rules doesn't necessarily imply that these variables are good discriminantes of the totality of the observations, once in most of the partition nodes only a part of the observations is considered. Thus, in a node other than the first only the remaining observations, not yet classified in a leaf node are considered. Even the first node, where all observations are considered, is different from a traditional discrimination approach. The reason is that a variable is chosen to discriminate between one group and the remaining ones not the totality of the groups simultaneously.

In steep three of the methodology used, regression analysis was used to forecast annual sales including information about the analogue group resulting from the classification. To determine the relative importance of the different types of variables used as predictors in the forecasting procedure, **dominance analysis** (Budescu, 1993) is considered adequate. The standardized regression coeficient is one of the more comomn measures, and can be interpreted as the expected change in the target associated with a change of one standard deviation in the predictor provided that all other predictors are held constant. In the special case of uncorrelated predictors a standardized coeficient is equivalent to the correlation between this predictor and the target, and in that way would provide an adequate measure of relative importance. Although a predictor whose coefficient is relativelly large will

presumably have a relative larger effect on the target, if the predictors are correlated, it may not be meaningful to think of the change in one predictor while all the others are keep constant because a change in one predictor will most likely result in a change in all predictors correlated with it. While conducted within a steepwise regression framework, dominance analysis is an alternative analytic strategy that assesses the relative importance of more than one set of variables to prediction using all combinations of predictor models (Azen and Budescu, 2003). The application of dominance analysis to the regressions used in the forecast steep of the suggested methodology result, once more, in the relevance of spatial analysis variables being only dominated by the "trade area".

In conclusion spatial analysis resulting variables are of special importance in site assessment studies and the delimitation method used for their calculation is relevant in this usefulness. Our suggestion consists in using several influence area delimitation methods simultaneously in a data analysis methodology framework.

References

Azen, R., Budescu, D., 2003. The dominance analysis approach for comparing predictors in multiple regression. Psycohological Methods 8 (2), 129-148.

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

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

Budescu, D.V., 1993. Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. Psychological Bulletin 114 (3), 542-551.

Cardoso, M.G.M.S., Moutinho, L., 2003. A logical type discriminant model for profiling a segment structure. Journal of Targeting, Measurement and Analysis for Marketing 12 (1), 27-41.

Cowen, D.J., Jensen, J.R., Shirley, W.L., Zhou, Y., Remington, K., 2000. Commercial real estate GIS site evaluation models: Interfaces to ArcView GIS. In: Proceedings of the 20º Annual ESRI International User Conference. ESRI online Library, pp. 140-145.

Crask, M.R., Perreault, W.D., 1977. Validation of discriminant analysis in marketing research. Journal of Marketing Research 11 (February), 60-64.

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.

Eurostat, 2001. Distributive trades in Europe. Office for Official Publications of the European Communities, Luxembourg.

Eurostat, 2003. European Business Facts and Figures, Part 5: Trade and tourism, data 1991-2001. Office for Official Publications of the European Communities, Luxembourg.

Gentle, J.E., 2002. Elements of Computational Statistics. Springer-Verlag, New York, USA.

Lattin, J., Carroll, J.D., Green, P.E., 2003. Analysing Multivariate Data. Duxbury, Pacific Grove, USA.

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., Cardoso, M.G.M.S., 2005. Clustering Supermarkets: The role of experts. Journal of Retailing and Consumer Services (in press).

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.

Mendes, A.B., Gonçalves, A.B., Oliveira, R.C., Matos, J., 2004. Sistema de Apoio à Decisão Espacial para localização de lojas de retalho: O problema das áreas de influência. In: Actas da 5ª CAPSI - Conferência da Associação Portuguesa de Sistemas de Informação. INESC, Lisboa, Portugal, pp. 1-11.

Quinlan, J.R., 1993. C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San Mateo, USA.

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.