Paper 492-2013

The Value of Neighborhood Information in Prospect Selection Models: Investigating the Optimal Level of Granularity

Philippe Baecke
Area Marketing, Vlerick Business School, Belgium

Dirk Van den Poel
Faculty of Economics and Business Administration, Department of Marketing, Ghent University, Belgium

ABSTRACT

Within analytical customer relationship management (CRM), customer acquisition models suffer the most from a lack of data quality because the information of potential customers is mostly limited to socio-demographic and lifestyle variables obtained from external data vendors. Particularly in this situation, taking advantage of the spatial correlation between customers can improve the predictive performance of these models. This study compares the predictive performance of an autoregressive and hierarchical technique in an application that identifies potential new customers for 25 products and brands. In addition, this study shows that the predictive improvement can vary significantly depending on the granularity level on which the neighborhoods are composed. Therefore, a model is introduced that simultaneously incorporates multiple levels of granularity resulting in even more accurate predictions.

1. INTRODUCTION

As markets become increasingly saturated and highly competitive, companies have shifted their marketing strategies from transactional marketing to relationship marketing (Coussement et al. , 2010; Pai & Tu, 2011). This is reflected in an explosion of interest in customer relationship management (CRM) by both academics and business practitioners (Ngai et al., 2009). Due to the information revolution and the drop in costs of data warehousing, many companies have collected a vast amount of socio-demographic and transactional data of their customers. In addition, computer power is increasing rapidly and data mining techniques are used to exploit this data in an optimal manner (Hosseini et al., 2010; Kamakura et al., 2005). This has resulted in the development of a wide range of software tools which enable companies to transform the collected data into useful information for marketing decision makers.

Besides the data mining technique, the success of a CRM model also depends on the quality of the information used as input for the model (Baecke & Van den Poel, 2011). Traditional CRM models often ignore neighborhood information and rely on the assumption of independent observations. This means that customers' purchasing behavior is totally unrelated to the behavior of others. However, in reality, customer preferences do not only depend on their own characteristics, but are often also related to the behavior of other customers in their neighborhood. Using neighborhood information to incorporate spatial autocorrelation in the model can solve this shortcoming and significantly improve the predictive performance of the model.

From all CRM fields, it is often most difficult to obtain good predictive results in the case of customer acquisition. This is because obtaining information from potential customers is not straightforward (Thorleuchter et al., 2012). As a result, in order to identify possible prospects, acquisition models are often estimated only based on a limited number of variables obtained from external data vendors (Baecke & Van den Poel, 2011). Especially in such a context where the availability of data is limited, incorporating neighborhood effects can be very valuable.

In academic literature, there are two important studies that specifically focus on the incorporation of spatial interdependence in order to improve customer identification, each using a different predictive technique. On the one hand, Yang & Allenby (2003) used an autoregressive approach to incorporate both geographic and demographic proximity between customers in a CRM model that predicts customers' preference for Japanese-made cars. That study indicated that geographic reference groups still have a larger impact than demographic reference groups. On the other hand, Steenburgh et al. (2003) used a hierarchical model to include a massively categorical variable, such as zip-codes, in order to improve the acquisition of new students at a private university. This paper contributes to previous literature by comparing the predictive performance of these two predictive techniques across multiple product categories.

Furthermore, this study can deliver interesting insights for a marketing decision maker. Currently, most research on spatial interdependence has been devoted to publicly consumed durable goods, such as automobiles (e.g. Yang & Allenby, 2003). This is because these highly visible products are more likely to be subject to social influence (Bearden & Etzel, 1982). However, until now, almost no attention has been given to the existence of neighborhood effects in less visible or less involving product categories. Besides applying spatial models on publicly consumed durable goods, this paper will also focus on privately consumed durable goods and consumer packaged goods.

Besides giving an overview across industries, this study will also focus on the optimal level of granularity on which neighborhood effects are optimally included in the model. Customers can often be clustered in neighborhoods at multiple levels (e.g. country, district, ward, etc.). In order to incorporate these neighborhood effects efficiently, the level of granularity should be carefully chosen. If the neighborhood is chosen too large, the spatial interdependence will fade away because the preferences of too many surrounding customers are taken into account that do not have any influence in reality. On the other hand, choosing neighborhoods that are too small can affect the stability of the measured influence and ignore the correlation with some customers that still have an influence. Based on data about the purchase of a Japanese car brand, this study will compare the relevance of taking neighborhood effects into account at different levels of granularity.

In order to facilitate the decision making about the optimal granularity level, a model is introduced that simultaneously incorporates multiple levels. Such a model is developed based on the assumption that multiple sources are responsible for the existence of autocorrelation between customers' purchasing behaviors (e.g. word of mouth, observational learning, homophily and other exogenous shocks) and each of these sources will have a different range in which interdependence exists. As a result, this model is able to incorporate spatial autocorrelation from several sources, each at their optimal granularity level.

The remainder of this paper is organized as follows. The data is described in section 2. Section 3 elaborates on the evaluation criterion and the predictive classification techniques used in this study are explained in section 4. The results are reported in Section 5 and Section 6 provides a discussion of these results in combination with a conclusion.

2. DATA DESCRIPTION AND PRODUCT CATEGORIES

This paper is based on data collected from one of the largest external data vendors in Belgium. Multiple socio-demographic and lifestyle variables are used as predictors to identify customers with a preference for a particular product or brand. An overview and description of these variables can be found in Table 1.

Next to the independent variables, also a discrete zip code variable is used to group customers into 589 mutually exclusive neighborhoods. Similar to the papers of Yang & Allenby (2003) and Steenburgh et al. (2003), spatial interdependence is assumed between customers living in the same neighborhood. This paper gives an overview for which products and brands spatial interdependence can be observed and investigates whether taking the spatial structure of the data into account can improve CRM predictions for prospect selection. Table 2 presents all products and brands examined in this study, divided into three main groups, namely public durable goods, private durable goods and consumer packaged goods. As shown in the last two columns of Table 2, which represent the number of observations and the number of events of each dependent variable, this study is based on a very large data sample.

In general, research on spatial interdependence and social influence is typically carried out on durable goods. For these products, neighborhood effects are more likely to be identified because they are purchased infrequently and relative expensive, resulting in a higher involvement of the customer. Besides involvement, also the visibility of the product could have an impact on the existence of interdependence between customers' purchasing decisions (Bearden & Etzel, 1982). Products for which the consumption is very visible will be more subject to reference group influence than privately consumed products. Therefore, durable goods are split into a publicly consumed and a privately consumed category. In the publicly consumed category five automobile brands, each brand originally coming from a different country, and five large clothing brands are examined. However in the privately consumed category, the focus will be on the purchase of five products, irrespective of the brand. This is based on Bearden & Etzel (1982) who illustrated that for publicly consumed durable goods, reference group influence mainly affects the brand choice decision, whereas for privately consumed goods the product choice decision will be mostly influenced.

Besides examining durable goods, this study will also explore the effect of incorporating spatial interdependence to identify customers of consumer packaged goods (CPGs). CPGs are typically low-involvement products with very low risk associated to the purchase. As a result, investigating the existence of spatial interdependence for these products has been ignored by literature for a long time. Since these products are frequently bought by everyone, almost no

differentiation would be measured in terms of purchasing behavior of the product itself. Therefore, in this category the focus will be on brand-choice influences. Hence, ten CPG brands are included in this research divided over two product categories (i.e. sodas and shampoos).

Variable name	Description
Socio-demographic variables:	
Age	The subject age divided over 14 age groups
Gender	The gender of the subject
Income	The income of the subject divided over 5 classes
Language	The language of the subject
Head_of_family	Whether the subject is head of the household
Pers_fam	The number persons in the household of the subject
Kids	The number of kids in the household of the subject divided over 4 age groups
Director	The subject is a self_employed earner, a director, a manager at a puplic limited company or a manager at a private limited company
Nb_household	The number of households in the building of the subject
Lifestyle variables:	
26 variables ranging from 0 to 1 indicating the interest of a subject into particular product categories: *Active sports, Cars, Cell phone, Cleaning products, Clothes, Consumer credits, Culture, Decoration, Extra insurance, Food and drinks, Grocery shopping, Holidays, Internet, Magazines, Multimedia, Multimedia equipment, Newspapers, Non-profit, No-risk investments, Omnium insurance, Risk investments, Passive sports, Pay-TV, Personal hygiene, Telephoning, Wellness*	

Table 1. Overview of independent variables

For each of the products and brands in Table 2, this study will investigate, based on two modeling techniques, whether neighborhood effects can be observed and whether these discovered effects are strong enough to improve a traditional customer acquisition model.

Besides this data, also information about the geographical location of the respondents is needed. For this, spatial variables are used provided by the external data vendor company that divides customers into mutually exclusive neighborhoods (e.g. zip-codes). Such variables can be obtained easily and, as a result, are frequently used for spatial analysis in marketing (Bradlow et al., 2005; Bell & Song, 2007; Steenburgh et al, 2003). These neighborhood indicators are often constructed on multiple levels of granularity (e.g. country, district, ward, etc.). Hence, the level on which the respondents are grouped can have an influence on the predicted performance of the model. Therefore, this study will investigate for one specific product (i.e. a Japanese car brand) a wide variety of granularity levels offered by the external data vendor. Table 3 presents the seven granularity levels examined in this study in combination with information about the number of neighborhoods at that level, the average number of respondents and the average number of owners (of a particular product) in each neighborhood.

Analysis based on a finer level of granularity will divide the respondents over more neighborhoods resulting in a smaller number of interdependent neighbors. At the finest level, an average of about 20 respondents is present in each neighborhood, which corresponds with an average of only 0.18 owners per neighborhood. This study will investigate which granularity level is optimal to incorporate customer interdependence using a generalized linear autologistic regression model.

		No. obs.	No. events
Public Durable Goods			
Automobiles	*Ford*	3143374	118192
	Toyota	3143374	85711
	Mercedes	3143374	57518
	Fiat	3143374	30759
	Volvo	3143374	26134
Clothes	*C&A*	617431	243297
	E5 Mode	617431	140613
	Zara	617431	100577
	Scapa	617431	44269
	Mango	617431	34856
Private Durable Goods			
Microwave		1348662	850068
Dish washing machine		1800293	690514
Surround system		954275	589288
Refrigerator with freezer		571372	344221
Espresso Machine		786511	121062
Consumer Packaged Goods			
Sodas	*Coca-Cola*	338735	114032
	Fanta	338735	61520
	Ice Tea	338735	54583
	Sprite	338735	41870
	Aquarius	338735	25570
Shampoos	*Dove*	342454	63626
	Elseve	342454	61845
	Fructis	342454	47003
	Pantene	342454	42560
	Head & Shoulders	342454	39237

Table 2. Overview of examined products and brands

Granularity level	Number of neighborhoods	Average number of respondents	Average number of owners
level 1	9	349281.78	3073.00
level 2	43	73105.49	643.19
level 3	589	5337.07	46.96
level 4	3092	1016.67	8.94
level 5	6738	466.54	4.10
level 6	19272	163.11	1.44
level 7	156089	20.14	0.18

Table 3. Overview of granularity levels

3. EVALUATION CRITERION

In order to be able to evaluate the predictive performance of each model the database is randomly split into a training and validation sample. The training sample, containing 70% of the observations, is used to estimate the parameter estimates. Afterwards, each model is validated on the remaining 30% of observations.

The area under the receiver operating characteristic curve (AUC) is used as evaluation metric of the classifiers (Hanley & Mcneil, 1982). The advantage of an AUC in comparison with other evaluation metrics, like the percent correctly classified (PCC), is that PCC is highly dependent on the chosen threshold that has to be determined to distinguish the predicted events from non-events. The calculation of the PCC is based on a ranking of customers according to their *a posteriori* probability of purchase. Depending on the number of prospects to target a cutoff value is chosen. All respondents with an *a posteriori* probability of purchase higher than the cutoff are classified as prospects. All respondents with a lower likelihood of purchase are labeled as non-prospects. This classification can be summarized in a confusion matrix, displayed in Table 4 (Morrison, 1969).

		Predicted status	
		Buyer	**Non-buyer**
True Value	**Buyer**	True Positive (TP)	False Negative (FN)
	Non-buyer	False Positive (FP)	True Negative (TN)

Table 4. Confusion matrix

Based on this matrix the percentage of correctly classified observations can be formulated as (Bradley, 1997):

$$PCC = \frac{TP+TN}{TP+ TN+ FP+FN}$$

Besides the PCC, the following meaningful measures can also be calculated:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN+FP}$$

Sensitivity represents the proportion of actual events that the model correctly predicts as events (i.e. the number of true positives divided by the total number of events). Specificity is defined as the proportion of non-events that are correctly identified (i.e. the number of true negatives divided by the total number of non-events). It is important to notice that all these measures give only an indication of the performance at the chosen cutoff. In reality, the chosen cutoff will vary depending on the context of the problem of the decision maker, hence an evaluation criterion independent of the chosen cutoff, such as the AUC, is preferred.

The receiver operating characteristic (ROC) curve is a two-dimensional graphical representation of sensitivity and one minus specificity for all possible cutoff values used (e.g. Fig. 1). The AUC measures the area under this curve and can be interpreted as the probability that a randomly chosen positive instance is correctly ranked higher than a randomly selected negative instance (Hanley & Mcneil, 1982). This again illustrates that this evaluation criterion is independent of the chosen threshold. As a result, this criterion is often used as evaluation metric for the predictive performance of CRM. The AUC measure can range from a lower limit of 0.5, if the predictions are random (corresponding with the diagonal in Fig. 1), to an upper limit of 1, if the model's predictions are perfect.

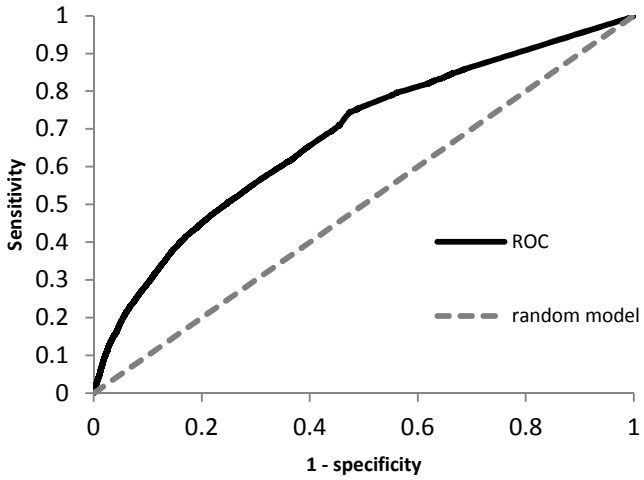


Fig. 1. AUC example

4. CLASSIFICATION TECHNIQUES

This study will try to predict whether or not a respondent has bought a particular brand or product. This results in a binary classification problem. This paragraph introduces three statistical techniques used throughout this study that are able to handle such problems. The traditional model is based on logistic regression techniques. Next, a spatial lag effect is added in an autologistic model that is able to incorporate spatial interdependence. Another way to incorporate this effect is by applying a hierarchical technique, such as a multilevel model. Once these models are built, probabilities can be estimated on a pool of potential new customers which helps to determine which of them has the highest chance to reply. Only addressing the customers with a high probability to purchase can already significantly improve the accuracy of a response model in direct marketing (Chen et al. , 2011). Consequently, a better performing prospect selection model can have a significant influence on a company's profit. Whereas a well-targeted mail can increase profits, an irrelevant mail will not only increase the marketing cost, but can also damage the image of a company on the long term (Kim et al., 2008).

4.1. LOGISTIC REGRESSION MODEL

Logistic regression is a well-known technique frequently used in traditional marketing applications (Bucklin & Gupta, 1992). An important benefit over other methods (e.g. neural networks) is its interpretability. It produces specific information about the size and direction of the effects of independent variables. Moreover, in terms of predictive performance and robustness, logistic regression can compete with more advanced data mining techniques (Levin & Zahavi,1998). Logistic regression belongs to the group of generalized linear models (GLM). GLMs adopt ordinary least square regression to other response variables, like dichotomous outcomes, by using a link function (McCullagh & Nelder, 1989). In logistic regression the parameters are estimated by maximizing the log-likelihood function. Including these estimates in the following formulae creates probabilities, ranging from 0 to 1, that can be used to rank customers in terms of their likelihood of purchase (Hosmer & Lemeshow, 2000).

$$P_i(y = 1 \mid all\ other\ variables) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

$$\eta_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni}$$

Whereby: P_i represents the *a posteriori* probability of purchase by customer i; X_{nj} represents the independent variables for customer i; β_0 represents the intercept; β_n represent the parameters to be estimated; n represents the number of independent variables.

Due to the high correlation between independent variables, it is possible that some variables, although significant in a univariate relationship, have little extra predictive value to add to the model. Hence, this study will include a backward selection technique that creates a subset of the original variables by eliminating variables that are either redundant or possess little additional predictive information.

The SAS code used to estimate such a logistic regression model is shown below:

```
PROC LOGISTIC DATA = inputtable_train OUTMODEL = parest_train;
MODEL &depvar. (EVENT='1') = &indepvars.
/SELECTION = backward SLSTAY = 0.001 STB;
OUTPUT OUT = predlog_train P = &depvar._pred;
ODS OUTPUT parameterestimates = log_paramest;
RUN;
```

PROC LOGISTIC is specifically designed for a logistic regression model. The procedure estimates parameters by means of maximum likelihood for a model with a binary dependent variable. The OUTMODEL option specifies the name of the data set containing sufficient information to score new data without having to refit the model. In the MODEL statement, a macro variable &depvar. refers to the dependent variable that needs to be explained using a macro list of independent variables. In the SELECTION option the backward selection method is specified based on a significance level of 0.001. The STB option adds standardized parameter estimates to the output. The OUTPUT OUT option creates a new dataset, called predlog_train, identical to the input dataset but with an extra column containing the predicted sales probabilities. The parameter estimates of the model are saved using the ODS OUTPUT statement.

Next, based on this model, prediction for the validation sample and the out-of-time test sample can be made using the following code:

```
PROC LOGISTIC INMODEL = parest_train;
SCORE DATA = inputtable_val OUT= predlog_val (rename = (p_1 =
&depvar._pred));
RUN;
```

This code uses the information from the parest_train dataset to make estimations based on the dataset defined in the DATA option. These predictions are saved in the dataset defined in the OUT option.

4.2. AUTOLOGISTIC MODEL

The autologistic model can be defined by means of the following equation (Besag, 1974, 1975):

$$P_i(y = 1 \mid all\ other\ variables) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

$$\eta_i = \beta_0 + \sum_{k=1}^{n} \beta_k X_{ki} + \rho \frac{\sum_{i \neq j} w_{ij} Y_j}{w_{ij}}$$

This equation is similar to a logistic regression model, but a spatial lag term is included that incorporates spatial interdependency. This spatial lag term is constructed based on an autoregressive coefficient ρ to be estimated for the spatially lagged dependent variable. This spatially lagged dependent variable is calculated using a weight matrix, which contains a one for customers living in the same neighborhood and a zero for every customer combination that lives in different neighborhoods (Anselin, 1988). By convention, self-influence is excluded such that diagonal elements equal zero. Next, this weight matrix is row standardized such that all row elements sum to one and multiplied with a vector containing the observed outcome variables. As such, the predicted behavior of a customer does not only depend on the customers' own characteristics but is also assisted by the behavior of neighboring customers.

The SAS code used to estimate such an autologistic model is very similar as a logistic model, as shown below:

```
PROC LOGISTIC DATA = inputtable_train OUTMODEL = parest_train;
MODEL &depvar. (EVENT='1') = &indepvars. spatial_lag_level1
/SELECTION = backward SLSTAY = 0.001 STB;
OUTPUT OUT = predlog_train P = &depvar._pred;
ODS OUTPUT parameterestimates = log_paramest;
RUN;
```

The only difference is that there is a spatial lag effect added to the independent variables. More specifically this a variable calculated based on a row standardized spatial weight matrix multiplied with a vector containing the observed outcome variables. Scoring the model on a validation sample is completely similar as a traditional logistic regression. However, important is that the spatial lag term is calculated only based on information of neighbors from the training sample. In this study, such a autologistic model will be estimated for each level of granularity.

Hence, at a coarse granularity level the amount of neighborhoods is small resulting in a high number of interdependent relationships included in the weight matrix. Consequently, the importance of the interdependent relationships of the customers that have an influence in reality could fade away because too much interdependence is assumed. As the granularity level becomes finer, the number of non-zero elements in the weight matrix will drop. However, if the level of granularity is too fine, the number of interdependent relationships could be too small, affecting the stability of the spatial lag effect. Therefore, this study will also investigate how the sample size of the dataset could influence the optimal granularity level.

Since the correlation among customers' purchasing behavior can have several origins (e.g. word of mouth, observational learning, homophily and other exogenous shocks), it is possible that this neighborhood effect can be divided into several sub-effects, each optimally estimated at a different granularity level. Hence, this paper will apply a model that incorporates spatial autocorrelation at multiple levels of granularity using the following formula:

$$P_i(y = 1 \,|all\ other\ variables) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

$$\eta_i = \beta_0 + \sum_{k=1}^{n} \beta_k X_{ki} + \sum_g \rho_g \frac{\sum_{i \neq j} w_{ijg} Y_j}{w_{ijg}}$$

In this model a separate autoregressive coefficient is estimated for each weight matrix constructed based on a different granularity level *g*. This should allow the model to incorporate each variety of spatial autocorrelation using its optimal measurement level, resulting in a more accurate predictive model.

4.3. MULTILEVEL MODEL

Another approach to include neighborhood effects in a binary predictive CRM model is by applying a multilevel model, also called a generalized linear mixed model (Breslow & Clayton, 1993; Wolfinger & O'Connell, 1993). This model does not include a spatial lag effect. Instead, it makes use of the hierarchical structure of the spatial data to incorporate interdependence of customers. Spatial models that specify a weight matrix, as previously explained, are based on 'Interaction Among Places' and state that objects that are close to each other are more related than distant objects, whereas multilevel models are related to 'Place Similarity' where the focus is more on hierarchy than on proximity (Anselin, 2002; Miller, 2004). In other words, these multilevel models state that objects in the same region are more related than objects in different regions. As a result, this model is only applicable when spatial data is used that divides customers into mutually exclusive neighborhoods (e.g. zip codes).

Assuming that data is available from *J* neighborhoods with a different number of customers n_j for each neighborhood, the complete formula of a multilevel model can be defined as follows (Hox, 2002):

$$P_i(y = 1 \,|all\ other\ variables) = \frac{exp(\eta_i)}{1 + exp(\eta_i)}$$

$$\eta_i = \beta_{0j} + \sum_{k=1}^{n} \beta_{kj} X_{ki}$$

This formula is related to a traditional logistic regression model, but it allows the intercept and slope coefficients, β_{0j} and β_{kj}, to vary across groups. These coefficients, often called random coefficients, have a distribution with a certain mean and variance that can be explained by *l* independent variables at the highest level Z_j, as follows:

$$\beta_{0j} = \gamma_{00} + \sum_{m=1}^{l} \gamma_{0m} Z_{mj} + u_{0j}$$

and

$$\beta_{kj} = \gamma_{k0} + \sum_{m=1}^{l} \gamma_{km} Z_{mj} + u_{1j}$$

The u-terms u_{0j} and u_{1j} represent the random residual errors at the highest level and are assumed to be independent from the residual errors e_{ij} at the lowest level and normally distributed with a mean of zero and a variance of $\sigma^2_{u_0}$ and $\sigma^2_{u_1}$ respectively. Since in this model errors are not assumed to correlate, a simple diagonal covariance matrix is used which models a different variance component for each random effect.

Because this model is used in a predictive context, containing a large amount of predictive variables, it is impossible to allow all slope coefficients to vary across groups. Certainly in combination with a large number of neighborhoods the model would become too complex, which may result in overfitting. Therefore, this model is simplified to a random intercept model, which can be written as (Baecke & Van Den Poel, 2010):

$$P_i(y = 1 \mid all\ other\ variables) = \frac{exp(\eta_i)}{1 + exp(\eta_i)}$$

$$\eta_i = \beta_{0j} + \sum_{k=1}^{n} \beta_k X_{ki}$$

where

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

In contrast to an autoregressive model in which a spatial lag effect is added, this model incorporates interdependence between the purchasing behaviors of customers in the same neighborhood by varying the intercepts for each neighborhood. As a result, customers living in the same neighborhood have a higher probability to show a similar purchasing behavior than customers living in different neighborhoods.

The SAS code used to estimate such a multilevel model is shown below:

```
PROC GLIMMIX DATA = inputtable METHOD= MSPL;
CLASS level1;
MODEL &depvar. (EVENT = '1') = &indepvars.
/DIST = binary LINK = logit SOLUTION;
RANDOM intercept / SUBJECT = level1 SOLUTION;
OUTPUT OUT = predtable pred(BLUP ILINK) = &depvar._pred;
ODS OUTPUT covparms = ml_covparamest parameterestimates = ml_paramest
solutionR = randeff;
RUN;
```

PROC GLIMMIX is a procedure recently developed by SAS in order to fit generalized linear mixed models. The input table contains one row for each respondent and includes a variable that divides each respondent into mutually exclusive neighborhoods (e.g. level 1). In this study, this model will be built multiple times, each time based on another level of granularity. This model is estimated using the maximum log-likelihood and the subject-specific

expansion principles (METHOD = MSPL). The class statement includes all variables that are categorical. Obviously, this neighborhood variable is included in the class statement since it is a categorical variable. Just like in PROC LOGISTIC, the MODEL statement includes a dependent variable that has to be modeled using a macro list of fixed effects. In the options of the MODEL statement it is defined that the distribution of the outcome variable is binary and a logit link function should be used for transformation. The significance of the effects can be evaluated using a t-test, which will be provided with the parameter estimates using the SOLUTION option. The RANDOM statement specifies that the intercept can vary across neighborhood levels. The OUTPUT OUT option creates a new dataset, called predtable, identical to the input dataset but with an extra column containing the predicted values based on the fixed and random effects (BLUP option), mapped onto the probability scale (ILINK option). The covariance parameter estimates and the solutions for fixed and random effects are saved using the ODS OUTPUT statement.

Unfortunately, the PROC GLIMMIX procedure does not provide a SCORE statement as in the PROC LOGISTIC procedure. As a result, an alternative approach has to be used to train the model on a training sample and score this on the validation sample. This can be easily accomplished by executing the following steps:

1. Replace the values of the dependent variable to a missing value (however, keep track of these real values in another variable: e.g. &depvar._copy)
2. Stack the training sample, which still has values for the dependent variable &depvar., and the validation sample, which only has missing value in the &depvar. variable.
3. This final table called inputtable is used as input for the PROC GLIMMIX procedure. By this the model will only be estimated on the observation which do not contain missing values in the dependent variable (i.e. training sample), however prediction scores will be calculated for all observations that have no missing values in the independent variables (i.e. both the training and validation sample).
4. In order to evaluate the predictive performance on the validation sample you will have to make use of the &depvar._copy variable, which still contains the real values of the dependent variable

5. RESULTS

In this chapter an overview of the results will be presented. In the first subsection a comparison is made across 25 products and brands in terms of predictive performance between a traditional logistic regression model and two techniques that are able to incorporate spatial interdependence. This comparison is only made on one level of granularity. However based on another dataset of a Japanese car brand, respondents can be divided into neighborhoods on multiple levels of granularity. One of the advantages of an autologistic model compared to a multilevel model is that this model can be extended so that it can easily incorporate multiple granularity levels simultaneously, whereas this is typically limited to only 3 levels in a multilevel model (Hox, 2002). Therefore, starting from subsection two, results are only based on an autologistic regression model. Subsection 2 compares the predictive improvement between models in which spatial interdependence is incorporated at different levels of granularity. Subsection 3, compares the predictive improvement of the best performing single level model, selected in in subsection 2, with a proposed model that is able to incorporate all granularity levels simultaneously.

5.1. MODEL COMPARISON

Table 5 demonstrates how neighborhood information can give extra value to a prospect selection model. This table compares for each product and brand the predictive performance in terms of AUC on the validation sample of a traditional logistic regression model, used as benchmark model, with an autologistic model and a multilevel model in which neighborhood effects are incorporated. In a comparison of the predictive performance of the models based on the non-parametric test of DeLong et al. (1988) using a 0.001 confidence interval, Table 5 shows that for all products and brands both spatial models perform significantly better than a traditional logistic regression model. This means that not only for public durable goods, but also for privately consumed durables and consumer packaged goods a significant improvement can be observed. When comparing both spatial models, the non-parametric test of DeLong et al. (1988) indicates that in 11 of the 25 cases the multilevel model significantly outperforms the autologistic model. Especially when the purchasing behavior of durable goods is modeled, the use of a multilevel model is preferred. Since the purchases of these goods are more influenced by neighborhood effects, the way how these influences are included on top of traditional variables will have a larger impact on the total predictive performance. Hence, for these durable goods the multilevel model is superior in even 10 out of the 15 cases.

		Benchmark Model	Autologistic Model[1]	Multilevel Model[2]
Public Durable Goods				
Automobiles	*Ford*	0.6350	0.6566	0.6568
	Toyota	0.6387	0.6577	0.6582
	Mercedes	0.7399	0.7439	0.7448*
	Fiat	0.6482	0.6656	0.6674*
	Volvo	0.6976	0.7041	0.7054
Clothes	*C&A*	0.6755	0.6894	0.6922*
	E5 Mode	0.6921	0.7125	0.7131*
	Zara	0.7800	0.7885	0.7893*
	Scapa	0.8194	0.8227	0.8242*
	Mango	0.8050	0.8120	0.8117
Private Durable Goods				
Microwave		0.6993	0.7024	0.7029*
Dish washing machine		0.7220	0.7247	0.7256*
Surround system		0.7144	0.7160	0.7167*
Refrigerator with freezer		0.5947	0.5982	0.5984
Espresso Machine		0.6577	0.6624	0.6634*
Consumer Packaged Goods				
Sodas	*Coca-Cola*	0.6230	0.6240	0.6244
	Fanta	0.6882	0.6901	0.6902
	Ice Tea	0.7210	0.7227	0.7234
	Sprite	0.6958	0.6978	0.6980
	Aquarius	0.7459	0.7484	0.7493*
Shampoos	*Dove*	0.6403	0.6422	0.6423
	Elseve	0.6342	0.6364	0.6371
	Fructis	0.6732	0.6752	0.6747
	Pantene	0.6472	0.6493	0.6498
	Head & Shoulders	0.6531	0.6557	0.6556

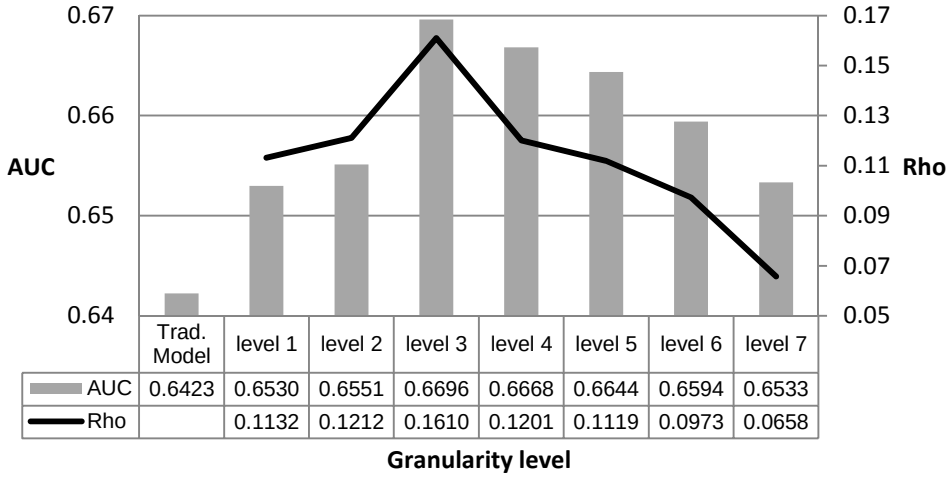
[1] All AUCs of the autologistic model differ significantly from the benchmark model on a 0.001 significance level
[2] All AUCs of the multilevel model differ significantly from the benchmark model on a 0.001 significance level
* Significant difference between autologistc and multilevel model on a 0.001 significance level

Table 5. Overview of the predictive performance in terms of AUC

5.2. SINGLE LEVEL AUTOLOGISTIC MODEL

In Fig. 2, the traditional prospect selection model and all “single level” spatial models are compared. This figure presents for each model the predictive performance on the validation sample in terms of AUC and the autoregressive coefficients estimated by the spatial models. All these performance criteria are based on models that try to identify prospects for a Japanese car brand.

The spatial autoregressive coefficients are positive and significantly different from zero in all autologistic regression models. This suggests the existence of interdependence at all levels of granularity. In other words, the average correlation between automobile preferences of respondents in the same neighborhood is higher than the average correlation between automobile preferences of respondents located in different neighborhoods. Comparing the AUC indicators of the spatial models with the benchmark traditional logistic regression model using the non-parametric test of DeLong et al. (1988), demonstrates that incorporating these neighborhood effects significantly improves the accuracy of the acquisition model.



	Trad. Model	level 1	level 2	level 3	level 4	level 5	level 6	level 7
AUC	0.6423	0.6530	0.6551	0.6696	0.6668	0.6644	0.6594	0.6533
Rho		0.1132	0.1212	0.1610	0.1201	0.1119	0.0973	0.0658

Fig. 2. Overview of the AUCs and the spatial autoregressive coefficients

However, the proportion of this predictive improvement heavily depends on the chosen granularity level. The optimal predictive performance in this study is achieved at granularity level 3. If the neighborhood level is too coarse, correlation is assumed between too many customers that do not influence each other in reality. On the other hand, a model based on a granularity level that is too fine could ignore interdependent relationships that exist in reality and affect the stability of the spatial lag effect because the number of customers in each neighborhood is too small. A similar evolution can be found in the spatial autoregressive coefficient (rho), which represents the existence of spatial interdependence in the model.

Comparing the predictive performance of a prospect selection model that incorporates neighborhood effects at the optimal granularity level with the benchmark traditional logistic regression model illustrates that taking spatial correlation into account heavily increases the AUC by 2.73 percentage points.

5.3. ALL LEVELS AUTOLOGISTIC MODEL

In Table 6, a comparison is made between the benchmark logistic regression model, the best performing spatial model at granularity level 3 and a model that simultaneously includes all granularity levels. This table gives an overview of the spatial autoregressive coefficients and the predictive performance of each model in terms of AUC.

This comparison proves the value of simultaneously including all granularity levels. Whereas in the first spatial model all neighborhood effects need to be captured in one spatial autoregressive coefficient, the second model makes it possible to estimate spatial correlation at several granularity levels. As a result, the spatial autoregressive coefficients are significant at five different neighborhood levels. Interdependence between customers' purchasing behavior is still best measured at level 3, but the model is also able to capture neighborhood effects on a coarser level 1 and several finer granularity levels (i.e. level 4, 5 and 7). The spatial autoregressive coefficients at level 2 and level 6 are not significant, implying that the spatial interdependencies measured by these two spatial lag effects are already covered by other spatial variables.

Variable	Stand. est. benchmark model	Stand. est. spatial model (level 3)	Stand. est. spatial model (all levels)
Spatial autoregressive coefficients (ρ):			
level 1			0.0412
level 3		0.1610	0.0935
level 4			0.0337
level 5			0.0299
level 7			0.0485
AUC:	**0.6423**	**0.6696**	**0.6783**

Table 6. Overview of the benchmark model, the spatial model at granularity level 3 and the spatial model including all granularity levels

Such a model is able to improve the AUC with an extra 0.87 percentage point compared to the best spatial model based on a single weight matrix which means a total improvement of 3.60 percentage points compared to a traditional CRM model. These results suggest that if the company has the resources to acquire multiple measurement levels of neighborhoods, it is advisable to simultaneously include them in a spatial CRM model in order to obtain even more accurate predictions.

6. DISCUSSION AND CONCLUSION

Within customer relationship management, correctly identifying potential new customers can be a hard task because the information available is mostly limited to socio-demographic and lifestyle variables attracted from an external data vendor (Baecke & Van den Poel, 2011). In this context, augmenting these acquisition models with spatial information could improve the identification of prospects. Traditional CRM models often assume that customers act independently of each other, whereas in reality, the behavior of customers could be spatially correlated. In this case, it is preferable to use models that take advantage of this information instead of treating this as nuisance in the error term. This study applies two models (i.e. an autologistic model and a multilevel model) to investigate for 25 products and brands, divided over three categories, whether neighborhood effects could be identified and to what extent incorporating this spatial correlation can improve the predictive performance of customer acquisition models.

In a first step, the predictive performance of both spatial models is compared with a traditional CRM model. This comparison showed that both models are able to significantly improve the identification of customers across all of the 25 products and brands investigated in this study. When the predictive performance of both spatial models are compared with each other, this study finds that especially for durable goods, which are more exposed to neighborhood effects, a multilevel model is often better able to incorporate this spatial interdependence on top traditionally uses socio-demographic and lifestyle variables.

Further, this study also provides interesting insights for a marketing decision maker. Based on this comparison, involvement and visibility of a product turns out to be most determining whether neighborhood effects exist for a particular product or brand. The predictive improvement that results from the incorporation of spatial information is in general the highest for public durable goods. Compared to publicly consumed durable goods, the added value is already much more limited for privately consumed durables. For the identification of purchasers of specific CPG brands this added value is even smaller and, although still significant, economically less relevant.

In addition, this study indicates that the marketing decision maker should carefully choose the granularity level on which the neighborhoods are composed because this can have an important impact on the model's accuracy. In this research, the best predictive performance was obtained at granularity level 3. Estimations based at coarser granularity levels include too much interdependence that does not exist in reality, affecting the validity of the model. Conversely, if the level of granularity becomes too fine, the number of observations and events in each neighborhood declines, which can affect the stability of the spatial lag effect. Further, correlation could be ignored with customers that still have an influence in reality.

This study also points out that the existence of neighborhood effects can have multiple origins, such as social influences, homophily, and exogenous shocks. As a result, the underlying interdependence can be divided into multiple parts, each optimally measured on a different level of granularity. This paper shows that a model that

simultaneously includes multiple granularity levels is able to outperform the best "single level" autologistic model Hence, if the marketing decision maker has sufficient recourses it is advisable to obtain data which divides customers into neighborhoods at multiple granularity levels.

In general a multilevel model performs slightly better than an autologistic model when spatial interdependence is incorporated on a single granularity level. This is especially the case when trying to identify prospects for durable goods. However, when data is available that group customers into neighborhoods based on multiple levels of granularity an autologistic model is preferred. In fact, this last model is better able to incorporate multiple levels simultaneously, whereas a multilevel model is typically limited to 3 levels (Hox, 2002).

ACKNOWLEDGEMENT

REFERENCES

- Anselin, L. (1988). *Spatial econometrics: methods and models*. Dordrecht: Kluwer.
- Anselin, L. (2002). Under the hood Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, *27*, 247-267.
- Baecke, P., & Van Den Poel, D. (2010). Improving Purchasing Behavior Predictions By Data Augmentation With Situational Variables. *International Journal of Information Technology & Decision Making*, *09*(06), 853-872.
- Baecke, P., & Van den Poel, D. (2011). Data augmentation by predicting spending pleasure using commercially available external data. *Journal of Intelligent Information Systems*, *36*(3), 367-383.
- Bearden, W. O., & Etzel, M. J. (1982). Reference Group Influence on Product and Brand Purchase Decisions. *Journal of Consumer Research*, *9*(2), 183-194.
- Bell, D. R., & Song, S. (2007). Neighborhood effects and trial on the internet: Evidence from online grocery retailing. *Quantitative Marketing and Economics*, *5*(4), 361-400.
- Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of Royal Statistical Society, Series B (Methodological)*, *36*(2), 192-236.
- Besag, J. (1975). Statistical Analysis of Non-lattice Data. *Journal of Royal Statistical Society, Series D (The Statistician)*, *24*(3), 179-195.
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recogniti*on, 7, 1145-1159.
- Bradlow, E. T., Russell, G. J., & Bell, D. R. (2005). Spatial Models in Marketing. *Marketing Letters*, *16*(3-4), 267-278.
- Breslow, N. E., & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *88*(421), 9-25.
- Bucklin, R. E., & Gupta, S. (1992). Brand choice, purchase incidence and segmentation: An integrated modeling approach. *Journal of Marketing Research,* 29, 201–215.
- Chen, W.-C., Hsu, C.-C., & Hsu, J.-N. (2011). Optimal selection of potential customer range through the union sequential pattern by using a response model. *Expert Systems with Applications*, *38*(6), 7451-7461.
- Coussement, K., Benoit, D. F., & Van den Poel, D. (2010). Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert Systems with Applications*, *37*(3), 2132-2143.
- DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837-845.
- Hanley, J. A., & Mcneil, B. J. (1982). The meaning and use of area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29-36.
- Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression (second edition).*New York: John Wiley & Sons.
- Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, *37*(7), 5259-5264.
- Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*. New York: Taylor & Francis Group.
- Kim, D., Lee, H., & Cho, S. (2008). Response modeling with support vector regression. *Expert Systems with Applications*, *34*(2), 1102-1108.

- Levin, N., & Zahavi, J. (1998). Continuous predictive modeling: A comparative analysis. *Journal of Interactive Marketing,* 12, 5–22.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models (second edition)*. London: Chapman & Hall.
- Miller, H. J. (2004). Tobler ' s First Law and Spatial Analysis. *Annals of the Association of American Geographers*, *94*(2), 284-289.
- Morrison, D. G. (1969). On the interpretation of discriminant analysis. *Journal of Marketing Research,* 6, 156-163.
- Ngai, E., Xiu, L., & Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592-2602.
- Pai, J.-C., & Tu, F.-M. (2011). The acceptance and use of customer relationship management (CRM) systems: An empirical study of distribution service industry in Taiwan. *Expert Systems with Applications*, *38*(1), 579-584.
- Steenburgh, T. J., Ainslie, A., & Hans, P. (2003). Massively Categorical Variables: Revealing the Inforraation in Zip Codes. *Marketing Science*, *22*(1), 40-57.
- Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2012). Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. *Expert Systems with Applications*, *39*(3), 2597-2605.
- Yang, S., & Allenby, G. M. (2003). Modeling Interdependent Preferences. *Journal of Marketing Research*, *40*(3), 282-294.
- Wolfinger, R., & O'Connell, M. (1993). Generalized Linear Mixed Models: A Pseudo-Likelihood Approach. *Journal of Statistical Computation and Simulation*, *4*(3-4), 233-243.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Philippe Baecke
Area Marketing
Vlerick Business School
Reep 1
B-9000 Ghent
Belgium
E-mail : Philippe.Baecke@vlerick.com

Dirk Van den Poel
Faculty of Economics and Business Administration
Department of Marketing
Ghent University
Tweekerkenstraat 2
B-9000 Ghent
Belgium
E-mail : Dirk.VandenPoel@ugent.be
Website : http://www.crm.UGent.be