UNIVERSITEIT
GENT

**FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE**

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Including Spatial Interdependence in Customer Acquisition Models: a Cross-Category Comparison

Philippe Baecke [1]

Dirk Van den Poel [2]

May 2012

2012/788

[1] PhD candidate

[2] Corresponding author: Professor of Marketing Modeling, Ghent University, Tweekerkenstraat 2, B-9000 Gent, http://www.crm.UGent.be

D/2012/7012/21

Including Spatial Interdependence in Customer Acquisition Models: a Cross-Category Comparison

Philippe Baecke and Dirk Van den Poel*

Ghent University, Faculty of Economics and Business Administration, Department of Marketing,
Tweekerkenstraat 2, B-9000 Ghent, 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 an autoregressive and hierarchical technique that both are able to incorporate spatial information in a model that can be applied on a large dataset, typical for CRM. Predictive performances of these models are compared in an application that identifies potential new customers for 25 products and brands. The results show that when a discrete spatial variable is used to group customers into mutually exclusive neighborhoods, a multilevel model performs at least as well as, and for a large number of durable goods even significantly better than a more often used autologistic model. Further, this application provides interesting insights for marketing decision makers. It indicates that especially for publicly consumed durable goods neighborhood effects can be identified. Though, for the more exclusive brands, incorporating spatial information will not always result in major predictive improvements. For these luxury products, the high spatial interdependence is mainly caused by homophily in which the spatial variable is a substitute for absent socio-demographic and lifestyle variables. As a result, these neighborhood variables lose a lot of predictive value on top of a traditional acquisition model that typically is based on such non-transactional variables.

Keywords

Customer Intelligence • Data Mining • Autologistic Model • Multilevel Model • Neighborhood effects • Spatial Interdependence

***Corresponding author:** Dirk Van den Poel (dirk.vandenpoel@ugent.be)

Tel. : +32 9 264 89 80 / Fax.: + 32 9 264 42 79, http://www.crm.UGent.be

1. Introduction

As markets become increasingly saturated and highly competitive, companies have shifted their marketing strategies from transactional marketing to relationship marketing (Coussement, Benoit, & Van den Poel, 2010; Pai & Tu, 2011). In other words, companies are more focus on the acquisition of valuable customers, the development of these customers in order to make them even more profitable and the creation of a long-term relationships in order to improve customer loyalty and retention (Kamakura et al., 2005). This is also reflected in an explosion of interest in customer relationship management (CRM) by both academics and business practitioners (Ngai, Xiu, & Chau, 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, Maleki, & Gholamian, 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.

As a high quality database is the foundation of effective and efficient CRM, companies should invest in augmenting their database with extra valuable variables (Baecke & Van den Poel, 2011). In this context, several studies have proven that incorporating information about the geographic proximity between customers can be valuable in marketing (Bradlow, Russell, & Bell, 2005; Bronnenberg, 2005). This information can often be obtained at relatively low cost and could significantly improve the performance of a CRM model. Traditional CRM models assume that customers' decisions are unrelated to each other and only depend on the characteristics of the particular customer, whereas in reality, preferences are often also influenced by the purchasing behavior of other customers and their recommendations (Arndt, 1967). Besides this, the geographical location can also act as a proxy for socio-demographic information because agents with similar characteristics and tastes have the tendency to group together (Mcpherson, Smith-lovin, & Cook, 2001). As a result of this principle, called homophily, customers within the same neighborhood are often more homogeneous in terms of socio-demographic characteristics.

Although several studies have proven the existence of spatial interdependence between the purchasing behaviors of customers (Bell & Song, 2007; Bradlow et al., 2005; Bronnenberg, 2005; Grinblatt, Keloharju, & Ika, 2008), the incorporation of spatial information in a predictive CRM context is limited. To the best of our knowledge, only two studies have incorporated 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 to predict 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, Ainslie, & Hans (2003) used a hierarchical model to include a massively categorical variable like zip-codes in the model in order to improve the acquisition of new students at a private university. Though, also these two studies have some limitations. Firstly, until now, both techniques have never been compared in terms of predictive performance which makes it difficult for a marketing decision maker to choose the most appropriate technique. Secondly, due to the complexity of the spatial models, both studies are based on a small number of observations and predictive variables which does not match with current CRM applications. Thirdly, these studies were only based on one product or one university. Therefore, no real conclusion can be drawn about the applicability of these models on other product categories.

This paper contributes to these previous studies by investigating, using both an autoregressive and a hierarchical approach, how the incorporation of spatial interdependence can improve a CRM model. More specifically, this study will try to improve traditional customer acquisition models across multiple brands and products. 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, Van den Poel, & Prinzie, 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 in which the availability of data is limited, incorporating neighborhood effects can be very valuable. The purpose of the acquisition model in this study is to predict whether or not a respondent has bought a particular brand or product. These probabilities can then be estimated on a pool of potential new customers in order to determine which of them has the highest chance to reply. Only addressing the customers with a high probability to purchase will already significantly improve the accuracy of a response model in direct marketing (Chen, Hsu, & Hsu, 2011). This is important because the performance of a customer acquisition 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, Lee, & Cho, 2008).

Besides, comparing two spatial techniques across multiple products and brands, another contribution of this study is the quality and quantity of the data. Table 1 illustrates that compared to previous literature this paper is based on a larger and more realistic data sample. This is necessary since this study wants to investigate the added value of spatial information on top of the data traditionally used for customer identification. Hence, if only a small number of predictive variables were included, spatial information could easily become a significant predictor because it would act as a proxy for important missing variables.

INSERT TABLE 1 OVER HERE

Furthermore, the application in which the effect of spatial interdependence across multiple products and brands are compared 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. Grinblatt et al., 2008; 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 paid to the existence of neighborhood effects in less visible or less involving product categories. Therefore, besides applying spatial models on publicly consumed durable goods, this paper will also focus on privately consumed durable goods and consumer packaged goods.

The remainder of this paper is organized as follows. Section 2 will give an overview of all products and brands that will be examined in this study. In Section 3 the methodology is presented in which the two predictive models and the evaluation criteria are described. The results are reported in Section 4 and Section 5 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 will be used as predictors to identify customers with a preference for a particular product or brand. An overview and description of these socio-demographic and lifestyle variables can be found in Table 2.

INSERT TABLE 2 OVER HERE

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 will give an overview for which products and brands spatial interdependence can be observed and will investigate whether taking the spatial structure of the data into account can improve CRM predictions for customer acquisition. Table 3 presents all products and brands examined in this study, divided into 3 main groups, namely public durable goods, private durable goods and consumer packaged goods. As shown in the last two columns of table 3, 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.

INSERT TABLE 3 OVER HERE

In general, research on spatial interdependence and social influence is typically carried out on durable goods, such as automobiles (e.g. Grinblatt et al., 2008; Yang & Allenby, 2003). 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 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, 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. In each of the two durable goods categories a range of both luxury (e.g. “Mercedes”,”Volvo”, “Scapa”, “Espresso Machine”) and less luxury (e.g. “Toyota”, ”C&A”, “Refrigerator with freezer”) products and brands are included, because also this can have an impact on the amount of reference group influence.

Besides durable goods, this study will also examine 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. Only recently, two studies have discovered that for the purchase of CPGs also interdependence can exist. Kuenzel & Musters (2007) showed for low involvement products that some specific reference groups, such as close family or friend, can influence each other’s purchasing behavior. Although no influence was discovered by neighbors, this study will verify this based on real behavioral data instead of questionnaires. Also Du & Kamakura (2011) detected that customers who purchased a newly introduced CPG can influence the adoption decision of neighboring customers. Although these contagion effects were mostly temporally measured during the introduction of a new CPG, this paper will investigate whether neighborhood effects can also be detected for more established CPG brands in order to improve CRM models. Since these products are frequently bought by everyone, almost no differentiation would be measured in terms of product purchasing behavior. Therefore, in this category the focus will also be on brand-choice influences. Hence, ten CPG brands are included in this research divided over two product categories (i.e. sodas and shampoos).

For each of the products and brands in Table 3, this study will investigate, based on two modeling techniques, whether neighborhood effects can be measured and whether these discovered effects are strong enough to improve a traditional customer acquisition model.

3. Methodology

As previously mentioned, the purpose of an acquisition model is to predict whether or not a respondent has bought a particular brand or product. This binary classification problem is often solved in CRM by means of a logistic regression model, which will be used as benchmark model. This generalized linear model uses a logit link function to adopt ordinary least squares regression to a response variable with dichotomous outcomes (McCullagh & Nelder, 1989). The equation of this well-known model can be formulated as follows:

(1)

$$P(y = 1\ |all\ other\ variables) = \frac{exp(\eta)}{1 + exp(\eta)}$$

$$\eta = \beta_0 + \sum_{k=1}^{n} \beta_k X_{ki}$$

Whereby P represents the a posteriori probability that customer i is a buyer of a certain product; β_0 is the intercept; X_{ki} represents the independent variable k of customer *i; n* is the number of independent variables and β_k are the parameters that need to be estimated.

Several advantages have made this model a very popular technique in CRM. Unlike more complex predictive technique, this model is easily interpretable for managers. It provides information about the size and direction of the effect of each predictor (Hosmer & Lemeshow, 2000). Further, due to its popularity, this model is widely available in many statistical packages, providing quick and robust results (Neslin et al., 2006).

Despite these advantages, an important assumption of this traditional model is that customers are assumed to act independently of other individuals. However, in reality, a customers' behavior is often influenced by the behavior and recommendation of others. Several authors already recognized that agents who are situated geographically close to each other have a higher correlating behavior (Bradlow et al., 2005; Bronnenberg, 2005). As a result, instead of treating this as nuisance in the error term, including this interdependence could improve CRM prediction.

For this end, various techniques are discussed in the literature. In most studies a spatial autoregressive model is used to capture spatial interdependence (Bell & Song, 2007; Bronnenberg & Mahajan, 2001; Yang & Allenby, 2003). Such models create a spatial weight matrix to include the behavior of surrounding agents to assist in predicting the behavior of a particular customer. Although, when a spatial variable is used that divides customers into mutually exclusive neighborhoods, such as zip codes, also a hierarchical model can incorporate spatial interdependence (Steenburgh et al., 2003).

This paper will focus on two models, closely related to the models used in the research previously described, namely an autologistic model and a multilevel model. By means of both models this study will examine for multiple brands and products whether neighborhood effects can be observed. Next, the predictive improvement of these models with regard to a traditional model will be calculated. In the next two sections, the methodology of both models will be discussed.

3.1 Autologistic Model

Autologistic models are often used to model the distribution of animal and plant species (Augustin, Mugglestone, & Buckland, 1996; He, Zhou, & Zhu, 2003). However, recently, the advantages of these models have also been recognized in the field of marketing (Moon & Russell, 2008). The autologistic model can be defined by means of the following equation (Besag, 1974; Besag, 1975):

(2)

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

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

This equation is similar to the one for a logistic regression model, but a spatial lag term is included to incorporate 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.

3.2 Multilevel Model

Another approach to include neighborhood effects in a binary predictive CRM model is by using 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 in order to incorporate interdependence of customers. Spatial models that specify the weight matrix as in Equation (2) 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 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). Multilevel models are widely used in social sciences (Courgeau & Baccaini, 1998; V. E. Lee & Bryk, 1989), however in marketing, only Steenburgh et al. (2003) used such model to include neighborhood effects during the acquisition process of students for a private university. 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):

(3)

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

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

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:

(4)

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

and

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

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 $\sigma_{u_0}^2$ and $\sigma_{u_1}^2$ respectively. Since in this models 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 this 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):

(5)

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

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

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 per 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.

3.3 Evaluation Criteria

In order to evaluate the predictive performance, the database is split randomly into a training sample and a validation sample for each product or brand. 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 predictive performance of each model will be expressed in terms of the area under the receiver operating characteristic curve (AUC), which is graphically presented by a two-dimensional representation of sensitivity (i.e. the true positive rate) and 1-specificity (i.e. the false positive rate) (Huang & Ling, 2005). Mathematically, AUC can be calculated using the following formula (Hand & Till, 2001):

$$\hat{A} = \frac{S_0 - n_0(n_0 + 1)/2}{n_{0\, n_1}} \tag{6}$$

Whereby n_0 and n_1 are the number of observations in the dataset belonging to respectively class 0 and class 1 and S_0 is the sum of the of the class 0 test points. This calculates the probability that a randomly chosen positive instance is correctly ranked higher than a randomly selected negative instance (Hanley & Mcneil, 1982). This probability will be close to 0.5 if predictions are random and close to 1 for perfect predictions.

An important advantage of AUC compared to other performance criteria, such as the percent correctly classified (PCC), is its independence of the chosen cut-off. The PCC gives the performance at only one cut-off level on which instances are predicted to be in class 0 or class 1, whereas the AUC gives an overall value based on all threshold values. Furthermore, Huang & Ling (2005) claimed that in general, AUC is statistically more consistent and more discriminating than accuracy.

4. Results

Before investigating whether predictive improvement can be achieved by including neighborhood effects in a CRM model, first the individual predictive effect of spatial interdependence will be examined. Therefore, an empty model is estimated without any independent variables. In other words, the autoregressive model and the multilevel model will only make use of respectively the spatial weight matrix in and the hierarchical structure of the data in order to classify customers into buyers and non-buyers. This should give an indication of the amount of neighborhood effects that exists for each product or brand. In Figure 1 and 2, the predictive performance of both empty models is presented for each product and brand, divided over three product categories.

INSERT FIGURE 1 & 2 OVER HERE

In a first step, the difference in predictive performance between an empty autologistic model and an empty multilevel model will be compared by means of the non-parametric test of DeLong, DeLong, & Clarke-Pearson (1988). Since the number of observations is quite high a strict significance level of 0.001 is applied. This test indicates that for an empty model, the AUCs of both techniques do not significantly differ from each other. In other words, both models are equivalently able to measure the existence of spatial interdependence across all products and brands examined in this study.

Secondly, these figures illustrate that the existence of neighborhood effects depends on both the involvement and the visibility of the product. For public durable goods, a significant amount of customers' purchasing behavior can already be predicted by taking only the interdependent behavior of customers into account. Clearly, this is less for privately consumed durable goods and the lowest for consumer packaged goods. Next to involvement and visibility, the exclusivity of the product or brand seems to have also an influence, however to a lesser extent, on the existence of neighborhood effects. This can be derived from the relative high predictive performance of the models that predict the purchase of “Scapa”, “Mercedes” and “Volvo”, which are more luxury brands, compared to the other brands in their category. Also in the private durable goods category, a more luxury product such as an “Espresso machine” is ranked higher than necessities, such as a “Refrigerator with freezer”.

After examining neighborhood effects individually, Table 4 demonstrates how these effects can give extra value to a customer acquisition 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.

INSERT TABLE 4 OVER HERE

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 4 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 as a result of the incorporation of spatial information in the models. When comparing both spatial models with each other, the results deviate from the comparison based on the empty spatial models. Although, the predictive performance between both spatial techniques is statistically similar for some product and brands, 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.

INSERT FIGURE 3 & 4 OVER HERE

The improvement of each model is graphically represented in Figure 3 and Figure 4. In general, these figures follow the same trend as Figure 1 and 2 in such a way that also in terms of predictive improvement including neighborhood effects is most beneficial for public durable goods. Although, very remarkable is that within this product category, the most exclusive brands (i.e. “Scapa”, “Mercedes” and “Volvo”) are not able to benefit as much as the other brands while these luxuries experience the most spatial interdependence (see Figure 1 and 2). These luxury brands are mostly bought by a smaller, more specific group of customers. As a result, prospects can already be better identified using only socio-demographic and lifestyle variables. This is demonstrated by the high predictive performance based on only a traditional model in Table 4. In other words, the high spatial interdependence measured for these luxury brands is mainly caused by homophily in which the neighborhood variable is a substitute for the absent socio-demographic and lifestyle variables. This is also proven by table 5, in which for both predictive models the spatial parameters are compared between an empty model and a full model that includes also socio-demographic and lifestyle variables. More particular, for an autologistic model the impact of spatial interdepence is measured through the standardized spatial autoregressive coefficient, while in a multilevel model this is measured through the intercept variance estimate. All the spatial parameters in this table are significantly different from zero on a 0.001 significance level.

INSERT TABLE 5 OVER HERE

This table shows that for the more exclusive brands (i.e. “Scapa”, “Mercedes” and “Volvo”) the added value of the neighborhood variable reduces significantly on top of a traditional model in both models, while such a large drop of the spatial parameter estimates cannot be observed for the other public durable goods. For these brands, which are bought by a general public, it is more difficult to identify prospects only based on socio-demographic and lifestyle variables, resulting in a relatively poor traditional customer acquisition model (see Table 4). In such models, incorporating neighborhood can be very valuable to improve the identification of potential customers.

Compared to public durable goods, the benefits of including spatial information is a lot smaller for privately consumed durable goods and, although still significant, very low for consumer packaged goods.

5. 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. However, 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, both models perform equivalently when only spatial information is used as a predictor. Though, 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. By only using the spatial interdependence between customers, purchasing behavior is best predictable for public durable goods, followed by privately consumed durable goods. Predictions are worst for consumer packaged goods, which are not only privately consumed but customers are generally also low involved in these products. Within each of the durable goods categories, it can be recognized that, next to involvement and visibility, also the exclusivity of the product has an influence on the amount of spatial interdependence. In other words, customers of more luxury product and brand (e.g. “Scapa”, “Mercedes”, “Volvo”, “Espresso machine”) are easier to be identified based on only spatial information. With these findings, this paper confirms based on a large behavioral data sample the surveyed result of Bearden & Etzel (1982) who found that publicly consumed luxuries are exposed to the most reference group influence.

However, remarkable is that although these luxuries experience the highest spatial interdependence, the model improvement is smaller than expected after the enhancement of a traditional customer acquisition model with spatial information. This is caused by the fact that these brands are often bought by a typical and more exclusive group of customers which are already easier to identify based on only socio-demographic and lifestyle variables. Further, the spatial variable can be a good proxy for these independent variables resulting in relatively high predictive performance of a model that is only based on spatial information. However, once this spatial variable is used in combination with socio-demographic and lifestyle variables, it loses a lot of his predictive power. In other words, although publicly consumed luxury durables are the most exposed to neighborhood effects, augmentation of a customer acquisition model with spatial information is more valuable for products for which customers are difficult to be identified, which is often the case for more general, less exclusive brands.

In comparison with publicly consumed durable goods, the added value of incorporating neighborhood effects in models to identify customers of privately consumed durable goods is already less. For the identification of purchasers of specific CPG brands this added value is even smaller and, although still significant, economically less relevant. These findings are in line with the findings of Kuenzel & Musters (2007). Based on surveyed data, these authors found that also for low involvement products social influence can affect the purchase decision. Though, this only exists between specific reference groups, such as close family or friend, but not between neighbors.

Based on 25 products and brands, this paper gives clear indications to marketing decision makers that spatial interdependence should not be neglected for certain types of goods. Instead of treating this as nuisance in the error term, taking advantage of this phenomenon can significantly improve a customer acquisition model. However, in order to generalize these findings, future research should examine even more product and brands. Besides this, it would be interesting to investigate whether incorporating spatial interdependence could also improve other CRM models, such as cross-sell, up-sell or churn models, which also includes transactional variables next to socio demographic and lifestyle variables.

Acknowledgement

Both authors acknowledge the IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy).

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.

Arndt, J. (1967). Role of Product-Related Conversations in the Diffusion of a New Product. *Journal of Marketing Research*, *4*(3), 291-295.

Augustin, N., Mugglestone, M., & Buckland, S. (1996). An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, *33*(2), 339-347.

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.

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.

Bronnenberg, B. J. (2005). Spatial models in marketing research and practice. *Applied Stochastic Models in Business and Industry*, *21*(4-5), 335-343.

Bronnenberg, B. J., & Mahajan, V. (2001). Unobserved Retailer Behavior in Multimarket Data: Joint Spatial Dependence in Market Shares and Promotion Variables. *Marketing Science*, *20*(3), 284-299.

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.

Courgeau, D., & Baccaini, B. (1998). Multilevel analysis in the social sciences. *Population: An English selection*, *10*(1), 39-71.

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.

Du, R. E. X. Y., & Kamakura, W. A. (2011). Measuring Contagion in the Diffusion of Consumer Packaged Goods. *Journal of Marketing Research, 48*(1), 28-47.

Grinblatt, M., Keloharju, M., & Ika, S. (2008). Social Influence and Consumption: Evidence from the Automobile Purchases of Neighbors. *The review of Economics and Statistics*, *90*(4), 735-753.

Hand, D. J., & Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, *45*(2), 171-186.

Hanley, J. A., & Mcneil, B. J. (1982). The meaning and use of area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29-36.

He, F., Zhou, J., & Zhu, H. (2003). Autologistic regression model for the distribution of vegetation. *Journal of Agricultural, Biological, and Environmental Statistics*, *8*(2), 205-222.

Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression*. 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.

Huang, J., & Ling, C. X. (2005). Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*(3), 299-310.

Kamakura, W., Mela, C. F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., Naik, P., et al. (2005). Choice Models and Customer Relationship Management. *Marketing Letters*, *16*(3-4), 279-291.

Kim, D., Lee, H., & Cho, S. (2008). Response modeling with support vector regression. *Expert Systems with Applications*, *34*(2), 1102-1108.

Kuenzel, J., & Musters, P. (2007). Social interaction and low involvement products. *Journal of Business Research*, *60*(8), 876-883.

Lee, V. E., & Bryk, A. S. (1989). A Multilevel Model of the Social Distribution of High School Achievement. *Sociology of Education*, *62*(3), 172-192.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. London: Chapman & Hall.

Mcpherson, M., Smith-lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415-444.

Miller, H. J. (2004). Tobler ' s First Law and Spatial Analysis. *Annals of the Association of American Geographers*, *94*(2), 284-289.

Moon, S., & Russell, G. J. (2008). Predicting Product Purchase from Inferred Customer Similarity: An Autologistic Model Approach. *Management Science*, *54*(1), 71-82.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Detection Defection: Measuring of the Predictive Accuracy Understanding Models Churn Customer. *Journal of Marketing*, *43*(2), 204-211.

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.

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.

Yang, S., & Allenby, G. M. (2003). Modeling Interdependent Preferences. *Journal of Marketing Research*, *40*(3), 282-294.

Study	**Spatial technique**	**Dependent variable**	**Number of observations in training sample**	**Number of observations in validation sample**	**Number of zip-codes**	**Number of non-spatial variables**
Yang & Allenby (2003)	Hierarchical	Japanese car preference	666	191	122	6
Steenburgh et al. (2003)	Autoregressive	Enrollment of students	37,551	34,179	7279	9
This study	Hierarchical & Autoregressive	Purchasing behavior of 25 products and brands	between 237,114 and 2,200,361	between 101,621 and 943,013	589	35

Table 1 Comparison of data information between previous studies and this study

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
Kid	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 2 Overview of independent variables

		No. obs.	**No. events**
Public Durable Goods			
Automobiles	*Ford*	3143374	118192
	Toyota	3143374	85711
	Mercedes	3143374	57518
	Fiat	3143374	30759
Clothes	*Volvo*	3143374	26134
	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 3 Overview of examined products and brands

		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*
Clothes	*Volvo*	0.6976	0.7041	0.7054
	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

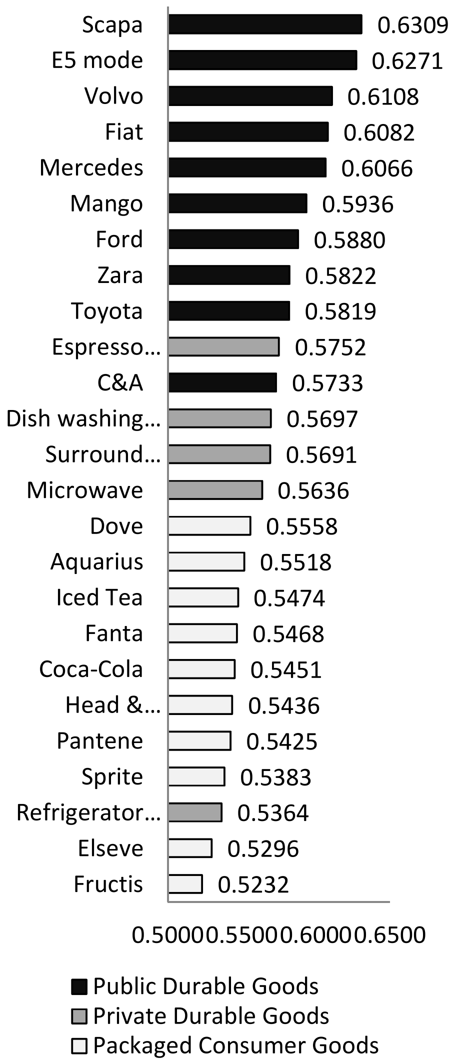
[2] All AUCs of the multilevel model differ significantly from the benchmark model on a 0.001 significance level

* Significant difference between autologistc and multilevel model on a 0.001 significance level

Table 4 Overview of the predictive performance in terms of AUC

		Autoregressive coefficient of autologistic model		Intercept Variance of multilevel model	
		Empty model	**Full model**	**Empty model**	**Full model**
Public Durable Goods					
Automobiles	*Ford*	0.1558	0.1528	0.1429	0.1259
	Toyota	0.1471	0.1436	0.1211	0.1267
	Mercedes	0.1944	0.1263	0.1840	0.0412
	Fiat	0.1863	0.1678	0.1840	0.1352
Clothes	*Volvo*	0.1973	0.1343	0.2147	0.0713
	C&A	0.1535	0.1547	0.0835	0.0974
	E5 Mode	0.2532	0.2413	0.1865	0.1286
	Zara	0.1638	0.1701	0.1113	0.1136
	Scapa	0.2629	0.1895	0.2553	0.1050
	Mango	0.1770	0.1744	0.1409	0.1119
Private Durable Goods					
Microwave		0.1276	0.1146	0.0597	0.0259
Dish washing machine		0.1408	0.0877	0.0622	0.0311
Surround system		0.1443	0.1023	0.0814	0.0246
Refrigerator with freezer		0.0748	0.0553	0.0246	0.0142
Espresso Machine		0.1470	0.1076	0.0921	0.0407
Consumer Packaged Goods					
Sodas	*Coca-Cola*	0.0909	0.0605	0.0372	0.0135
	Fanta	0.0918	0.0619	0.0444	0.0202
	Ice Tea	0.0966	0.0687	0.0503	0.0293
	Sprite	0.0772	0.0636	0.0362	0.0286
	Aquarius	0.1008	0.0820	0.0640	0.0476
Shampoos	*Dove*	0.1085	0.0727	0.0577	0.0199
	Elseve	0.0619	0.0500	0.0241	0.0161
	Fructis	0.0501	0.0407	0.0188	0.0163
	Pantene	0.0841	0.0574	0.0449	0.0188
	Head & Shoulders	0.0891	0.0600	0.0468	0.0218

Table 5 Overview of spatial parameters



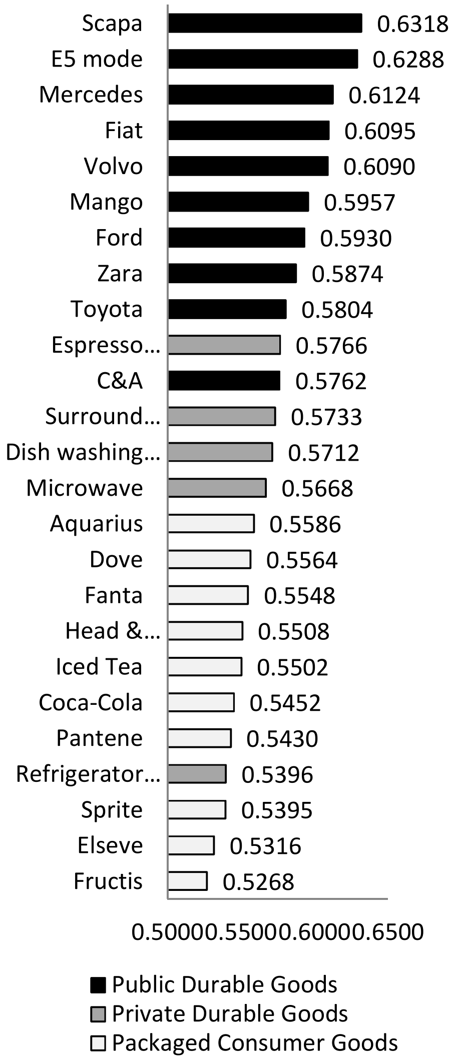


Fig 1 AUCs of an empty autologistic model

Fig 2 AUCs of an empty multilevel model

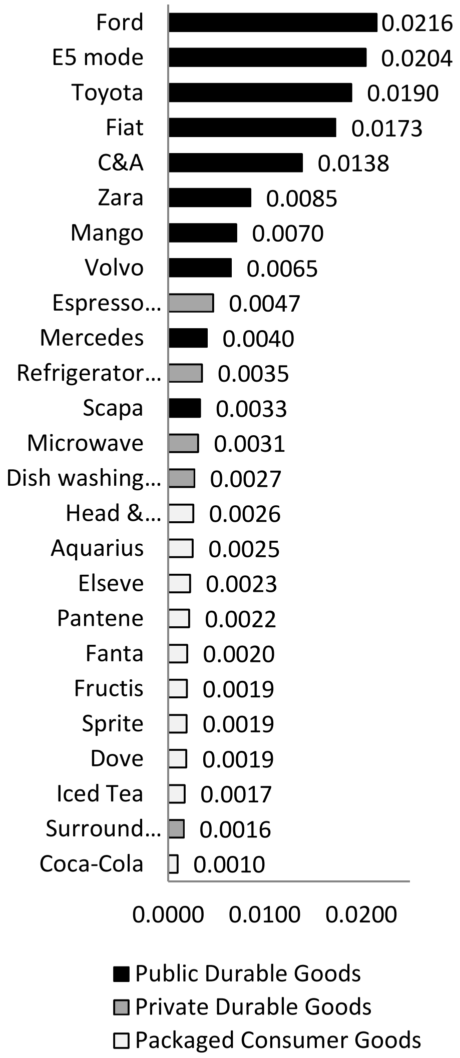


Fig 3 Predictive improvement of an autologistic model

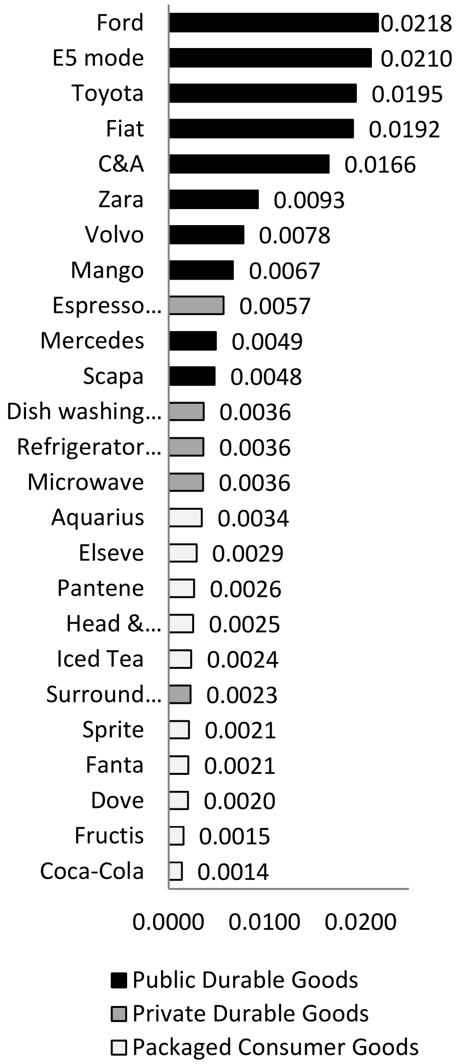


Fig 4 Predictive improvement of a multilevel model