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

Improving Customer Acquisition Models by Incorporating Spatial Autocorrelation at Different Levels of Granularity

Philippe Baecke [1]

Dirk Van den Poel [2]

October 2012

2012/819

[1] PhD, Ghent University, Faculty of Economics and Business Administration, Department of Marketing
[2] Corresponding author: Professor of Marketing Modeling, Ghent University, Tweekerkenstraat 2, B-9000 Gent, http://www.crm.UGent.be

D/2012/7012/52

Improving Customer Acquisition Models by Incorporating Spatial Autocorrelation at Different Levels of Granularity

Philippe Baecke[1] and Dirk Van den Poel[1]

[1]Ghent University, Faculty of Economics and Business Administration, Department of Marketing, Tweekerkenstraat 2, B-9000 Ghent, Belgium.

Abstract. Traditional CRM models often ignore the correlation that could exist among the purchasing behavior of surrounding prospects. Hence, a generalized linear autologistic regression model can be used to capture this interdependence and improve the predictive performance of the model. In particular, customer acquisition models can benefit from this. These models often suffer from a lack of data quality due to the limited amount of information available about potential new customers. Based on a customer acquisition model of a Japanese automobile brand, this study shows that the extra value resulting from incorporating neighborhood effects can vary significantly depending on the granularity level on which the neighborhoods are composed. A model based on a granularity level that is too coarse or too fine will incorporate too much or too little interdependence resulting in a less than optimal predictive improvement. Since neighborhood effects can have several sources (i.e. social influence, homophily and exogeneous shocks), this study suggests that the autocorrelation can be divided into several parts, each optimally measured at a different level of granularity. Therefore, a model is introduced that simultaneously incorporates multiple levels of granularity resulting in even more accurate predictions. Further, the effect of the sample size is examined. This showed that including spatial interdependence using finer levels of granularity is only useful when enough data is available to construct reliable spatial lag effects. As a result, extending a spatial model with multiple granularity levels becomes increasingly valuable when the data sample becomes larger.

Keywords: Customer Relationship Management (CRM); Predictive Analytics; Customer Intelligence; Marketing; Data Augmentation; Autoregressive Model; Automobile Industry

1 Introduction

Customer Relationship Management (CRM) has become an important topic in the field of marketing [1]. The technological development, the rise of the internet and declining costs for data warehousing and information processing have encouraged companies to collect data

about their customers and prospects [2]. CRM uses data mining techniques to convert this unstructured data into valuable information. This has resulted in the development of useful information technology tools to support marketing decision making and predict the effect of it [3,4].

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 [5]. 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.

Several studies have already proven that spatial statistics can produce interesting insights in marketing [6-12]. However, only a limited number of studies use spatial information to improve the accuracy of a predictive CRM model. In reference [13], customer interdependence was estimated based on geographic and demographic proximity. The study indicated that geographic reference groups are more important than demographic reference groups in determining individual automobile preferences. Reference [14] showed that taking zip-code information into account can significantly improve a model used for the attraction of new students by a private university. The focus of this research will also be only on physical geographic interdependence, but compared to previous literature, this study includes a high number of independent socio-demographic and lifestyle variables that are typically available at an external data vendor. This should prevent the predictive improvement to be caused by the absence of other important variables that can be easily obtained for customer acquisition models.

In this paper, neighborhood information is used to incorporate spatial autocorrelation in a customer acquisition model for a Japanese car brand. Reference [15] is the first paper that compared the value of incorporating spatial information in CRM models across multiple product categories. That study found that especially for publicly consumed durable goods, such as automobile brands, incorporating neighborhood effects can be very useful. Further, within CRM models,

customer acquisition models suffer the most from a lack of data quality. A company's customer database is typically single source in nature. The data collection is limited to the information a company retrieves from its own customers. As a result, for customer acquisition campaigns the company has to attract data from external data vendors. Nevertheless, these data still only contains a limited number of socio-demographic and lifestyle variables [16]. Especially in such situation, incorporating extra neighborhood information can improve the identification of potential customers.

In addition, an extra complexity is introduced that has been mostly ignored in previous literature. 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 reliability of the measured influence and ignore the correlation with some customers that still have an influence. 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 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.

Furthermore, this study will investigate how the size of the dataset can influence the predictive performance of the spatial models. These spatial models takes the the purchasing behavior of surrounding customers into account to assist in purchasing behavior predictions of a particular customer. At a finer level of granularity, customers are divided into more neighborhoods in which spatial interdependence is assumed. As a result, only closer neighbors, who are assumed to have a higher influence, are used to assist in the predictions. On the other hand though, this also results in fewer observations available to construct these spatial influences, which can eventually affect the reliability of

the spatial variables. Consequently, increasing the data sample should improve the incorporation of spatial interdependence calculated on finer granularity levels.
The remainder of this paper is organized as follows. Section 2 will elaborate on several sources that are responsible for the existence of spatial interdependence in CRM models. The methodology is described in Section 3, consisting of the data description, the generalized linear autologistic regression model and the evaluation criterion used in this study. The results are reported in Section 4 and Section 5 provides a discussion of these results in combination with a conclusion.

2 Origins of spatial interdependence

In this study neighborhood effects are defined as the existence of correlating purchasing behavior among geographically closely located customers. Based on previous literature, three concepts can be distinguished that are responsible for the existence of this spatial interdependence, namely social influence, homophily and exogenous shocks. The focus of this study is not to disentangle the effect of these three concepts, but to simultaneously take all these effects into account in order to obtain more accurate CRM models.
In the following sections these concepts are described, illustrating that the spatial autocorrelation caused by each effect may be optimally measured at different granularity levels. Hence, the added value of incorporating interdependence in a customer acquisition model can differ significantly depending on the granularity level that is used to compose the neighborhoods. Furthermore, a generalized linear autologistic regression model that allows dividing the spatial autocorrelation over multiple granularity levels can improve predictions even more.

2.1 Social influence

The power of social influences in marketing has been known for some time [17]. Customers do not live in an isolated environment where decisions are made in a purely rational way. Instead, product preferences and purchasing decisions are often influenced by positive and negative recommendations of other individuals. Word of mouth (WOM) can have an important impact on a customer's decision because this information is perceived as highly credible [18]. Due to its

non-commercial nature this information is processed with less skepticism than advertising or promotion. Although the emergence of online word of mouth should not be ignored, the majority of word of mouth conversations still take place in face-to-face interpersonal settings. More specifically, Reference [19] and [20] show that still 76% to 80% of the WOM conversations occur face-to-face, while only about 10% are online. Further, it can be assumed that people who live in the same neighborhood will have more correlated purchasing behavior, as living closer together provides more opportunities for interaction and communication. This has also been supported by reference [12] in which spatial proximity is used as proxy for WOM to investigate contagion in new product adoption. As a result, geographic proximity can still be considered as an important indication of social influence. Although online product recommendations will also have an influence on the purchasing behavior of the customers, already a large part of this social influence can be taken into account by using geographical information. In addition, spatial variables are ideal for data augmentation applications since these can be easily collected for a large number of customers.

Actually, customers do not even have to interact to affect each other. Observing the purchasing decisions of others can be enough to influence an individual's purchasing decision [21]. In other words, besides WOM, observational learning (OL) is a second important social influence that can be responsible for spatial autocorrelation in a CRM model. Neighboring customers buy similar products and brands not only because they want to match the social standard of the neighborhood, but also because they may be more confident about the quality if they recognize that many people bought the product or brand. Although WOM contains more information because it makes it possible to clarify an opinion or recommendation, the information from OL might be perceived as more credible because it reveals the real action of other consumers [22].

2.2 Homophily

Besides social influences, another explanation of the existence of interdependence between customers' purchasing behavior is homophily, also called endogenous group formation [23]. This concept is often referred to with the proverbial expression "Birds of a feather flock together" [24]. In other words, people with similar tastes and

characteristics tend to group together. Two types of homophily can be distinguished to explain the existence of sociospatial patterns, namely social homophily and structural homophily [25]. Social homophily means that people wish to live close to others with similar social characteristics. On the other hand, structural homophily refers to the fact that people with similar social characteristics may prefer similar physical attributes of neighborhoods. Due to these two types of homophily, residents with homogeneous characteristics will move to similar neighborhoods resulting in spatial patterns of socio-economic and demographic characteristics. This can explain spatially correlated purchasing behavior that is not created by the direct influence of one's behavior on another.

2.3 Exogenous shocks

A last cause of the existence of interdependence between customers is exogenous shocks. People of the same neighborhood may buy similar products or brands neither because they are influenced by each other nor because they have similar characteristics, but because they are subject to the same exogenous shock that exists in the neighborhood, such as promotional activities, the location of points of sales or even typical characteristics of the environment in the neighborhood.

3 Methodology

3.1 Data Description

Data is collected from one of the largest external data vendors in Belgium. This external data vendor possesses data about socio-demographics and lifestyle variables from more than 3 million respondents in Belgium. Furthermore, it provides information about automobile ownership in December 2007 of a Japanese automobile brand.

Table 1 gives an overview of all variables used throughout this study. The purpose of the proposed model is identifying respondents with a similar profile as current owners of the Japanese automobile brand, who can then be targeted using a marketing acquisition campaign. Hence, this customer identification model uses a binary variable as dependent variable, indicating whether the subject possesses the the Japanese car brand or not. A customer acquisition model often cannot rely on transactional information because company's customer databases are typically single source in nature and do not contain

information about non-customers [16]. Consequently, only a high number of socio-demographic and lifestyle predictors can be attracted from an external data vendor. The socio demographic variables contain variables that are traditionally included in a customer acquisition model. All categorical variables are split into n-1 dummies before they were included into the model. The lifestyle variables are variables created by the external data vendor indicating the interest of the respondent in a certain product category. These ratio summary variables were created based on multiple underlying questions and range from 0, if the respondent has totally no interest in the product category, to 1, if the respondent's interest is very high. Taking also these life style variables into account should prevent that the extra value resulting from incorporating neighborhood effects is caused by the absence of other important predictors that easily could be obtained from an external data vendor.

Table 1. Model Variables.

Variable name	Description
Dependent variable:	
Ownership	A binary variable indicating whether the subject possesses a particular Japanese automobile brand
Independent variables:	
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*

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 are easily attractable and as a result frequently used for spatial analysis in marketing [6,9,14]. These neighborhood indications 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 investigated a wide variety of granularity levels offered by the external data vendor. Table 2 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 in each neighborhood. Comparing the number of owners to the total number of observations indicates that the percentage of owners is relatively small (i.e. 0.88 %).This results from the facts that, firstly, not every respondent owns a car and, secondly, there exists a lot of competition in the automobile market resulting in a wide range of automobile brands to choose from.

Table 2. Overview of the granularity levels.

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

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, but also how the sample size can influence the power of these spatial variables.

3.2 Generalized Linear Autologistic Regression Model

A typical data mining technique used in CRM to solve a binary classification problem is a logistic regression model. This model is very popular in CRM because of its interpretability. Unlike other, more complex predictive techniques (e.g. neural networks), logistic regression is able to provide information about the size and direction of the effects of the independent variables [26,27].
A key assumption of this traditional model is that the behavior of one individual is independent of the behavior of another individual. Though, in reality, a customers' behavior is not only dependent of its own characteristics but is also influenced by the preferences of others. In traditional data mining techniques this interdependence is treated as nuisance in the error term. However, an autologistic regression model can be used to consider spatial autocorrelation explicitly in a predictive model for a binary variable. Originally, this model has been used in biological sciences [28-30], but recently it is also introduced in the field of marketing [10]. The generalized linear autologistic regression model in this study is a modified version of the general autologistic model introduced by Besang [31, 32]:

$$P(y = 1 \,|all\ other\ values) = \frac{exp(\eta)}{1 + exp(\eta)}. \qquad \textbf{(1)}$$
$$Where\ \ \eta = \beta_0 + X\beta_1 + \rho WY\ .$$

In this equation a logit link function is used to adopt the regression equation to a binomial outcome variable. Whereby Y is an n x 1 vector of the dependent variable; X is an n x k matrix containing the explanatory variables; the intercept is represented by β_0 and β_1 is a k x 1 vector of regression coefficients to be estimated.

This model includes also a spatial lag effect by means of the autoregressive coefficient ρ to be estimated for the spatially lagged dependent variables WY. These spatially lagged dependent variables are constructed based on a spatial weight matrix W.
The weight matrix is an important element in a generalized linear autologistic regression model and can be constructed in several ways. One way of creating the spatial weight matrix is based on the continuous distance between customers. Reference [13] for example assumed that geographical influence is an inverse function of geographical distance by using the following formula:

$$w_{ij} = \frac{1}{exp[d(i,j)]}. \quad (2)$$

In which d(i,j) represents the Euclidian distance calculated based on the latitude and longitude coordinates of the customers.
Within the field of marketing though, often a discrete spatial variable is used that divides customers into mutual exclusive neighborhoods (e.g. zip-codes) [6,9,14]. For such kind of variables the use of a contiguity matrix is more appropriate. Such matrix is constructed based on the relative positions of one customer to another. Since this study is focused on comparing discrete neighborhood variables, also a contiguity matrix will be used. This weight matrix is constructed based on an n x n matrix containing the elements w_{ij} indicating the interdependence between observation i (row) and j (column). Similar as in reference [13], w_{ij} will be set to one in a non-standardized weight matrix for customers living in the same neighborhood. By convention, self-influence is excluded such that diagonal elements w_{ij} equal zero. Next, this weight matrix is row-standardized using the following formula:

$$w_{ij}^{s} = \frac{w_{ij}}{\sum_j w_{ij}}. \quad (3)$$

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 fades 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 do be too small, affecting the reliability of the model. Therefore, this study will also investigate how the sample size of the dataset could influence the optimal granularity level.
Since the correlation existing between customers' purchasing behavior can have several origins (e.g. word of mouth and homophily), it is possible that this neighborhood effect can be divided into several sub effects, each optimally estimated on a different granularity level. Hence, this paper will apply a model that incorporates spatial autocorrelation at multiple levels of granularity using the following formula:

$$P(y = 1 \,|all\ other\ values) = \frac{exp(\eta)}{1 + exp(\eta)} \,. \qquad (4)$$
$$Where\ \eta = \beta_0 + X\beta_1 + \sum_g \rho_g W_g Y \,.$$

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.
Because this study is based on a high number of observations and variables, all model parameters are obtained using a maximum pseudolikelihood (MPL) estimation. Although more advanced techniques, such as Markov chain Monte Carlo (MCMC) [33] methods have been discussed in the literature, these techniques are not implemented because they are computationally infeasible for this large database. Furthermore, Reference [34] suggests that MPL estimates should be adequate when the spatial autoregressive coefficient is relatively small. In proportion to biological sciences, this is mostly the case in the field of marketing.
The model also includes a backward selection at a significance level of 0.0001 to eliminate redundant variables that do not add extra predictive

value. This should improve the comprehensibility of the model and decrease computational time and cost for scoring respondents [35].

3.3 Evaluation Criterion

In order to evaluate the predictive performance of the model, the database, containing more than 3 million observations, is randomly split into two parts. A training sample, consisting out of 70% of the observations, is used to estimate the model. Afterwards, this model is validated on the remaining 30% of observations. Several evaluation criteria, such as lift or PCC (percent correctly classified), suffer from the limitation that a cutoff value needs to be chosen to discriminate predicted events from non-events. The area under the receiver operating characteristic curve (AUC) solves this limitation by taking all possible thresholds into account [36]. The receiver operating characteristic (ROC) curve is a two-dimensional graphical representation of sensitivity (i.e. the number of true positives versus the total number of events) and one minus specificity (i.e. the number of true negatives versus the total number of non-events) for all possible cutoff values used. The area under this curve can range from a lower limit of 0.5 to an upper limit of 1. The closer this value is to one, the better the general accuracy of the model.

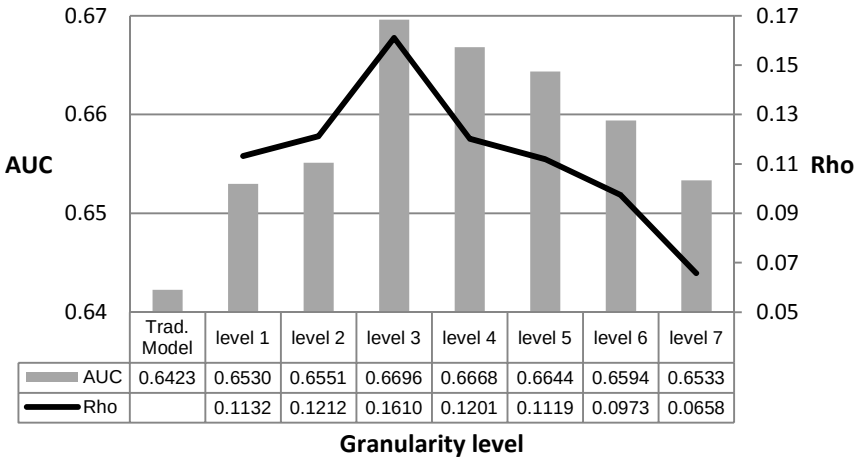
4 Results

In this chapter an overview of the results will be presented. In the first section a traditional logistic regression is compared with seven “single level” autologistic models that include spatial interdependence, each calculated based on a different level of granularity. Next, in the second section the best performing “single level” autologistic model is compared with a model that incorporates all levels of granularity simultaneously. In the last section, the effect of the sample size is examined on the predictive performance of the spatial models.

4.1 “Single level” autologistic model

In Fig.1, the traditional customer identification 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. These 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. [37], 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. 1. 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 reliability of the model 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 customer acquisition 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%. Although the differences between AUC can seem quite small, Reference [14] has illustrated that since such models are typically applied on a large number of prospects, even small differences in AUC can lead to large differences in terms of profitability. In other words, this improvement in predictive performance is not only statistically significant, but also economically relevant and should help marketing decision makers to improve their customer acquisition strategies.

4.2 "All levels" autologistic model

In Table 3, 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 all standardized parameter estimates of the socio-demographic and lifestyle variables that significantly influence automobile purchasing behavior at a 0.0001 significance level; the significant spatial autoregressive coefficients and the predictive performance of each model in terms of AUC.

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

Variable	Stand. est. benchmark model	Stand. est. spatial model (level 3)	Stand. est. spatial model (all levels)
Socio-demographic variables:			
Age group 18-21	-0.0548	-0.0586	-0.0592
Age group 22-25	-0.0241	-0.0256	-0.0264
Age group 31-35	-0.0292	-0.0260	-0.0268
Age group 36-40	-0.0359	-0.0345	-0.0356
Age group 61-65		0.0164	
Age group 66-70	0.0207	0.0235	0.0205
Age group 71-75	0.0165	0.0202	0.0175
Age group 76-80	0.0194	0.0230	0.0205
Is no director, self-employed earner or manager	0.0451	0.0437	0.0435
Manager at a private limited company	-0.0276	-0.0288	-0.0293
Number of persons in the household	-0.0669	-0.0628	-0.0662
Head of the household	-0.0614	-0.0547	-0.0553
Number of children younger than 5	-0.0222	-0.0232	-0.0228
Lifestyle variables:			
Cars	0.1265	0.1276	0.1262
Grocery shopping	0.1019	0.1003	0.1008
Magazines	0.0568	0.0542	0.0531
Clothes	-0.0541	-0.0633	-0.0590

Omnium insurance	-0.0439	-0.0374	-0.0355
Personal hygiene	0.0407	0.0467	0.0441
Passive sports	0.0375	0.0354	0.0380
Active sports	-0.0372	-0.0341	-0.0359
No risk investments	0.0369	0.0393	0.0397
Food and drinks	-0.0356	-0.0367	-0.0364
Cell phones	0.0299	0.0329	0.0329
Wellness	-0.0292	-0.0288	-0.0321
Consumer credit	0.0276	0.0282	0.0283
Newspapers	-0.0253	-0.0277	-0.0273
Culture	-0.0240	-0.0262	-0.0263
Telephoning	-0.0237		
Pay TV	0.0188		
Non-profit organizations		0.0201	0.0224
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**

Among the socio-demographic variables, age is a significant predictor. Older people are more likely to drive the Japanese automobile brand than younger people. Among the lifestyle variables, it is obvious that people who are more interested in cars are more likely to purchase the Japanese automobile brand. The parameter estimates of the three models do not differ a lot in size and direction. Except for one age group, i.e. age group 61-65, all the same socio demographic variables are significant. Considering the lifestyle variables, telephoning and pay TV turn out to be only significant in the benchmark model, whereas interest in non-profit organizations is only significant in the two spatial models.

More remarkable is that the spatial autoregressive coefficient already has the strongest influence of all parameters in the spatial model at granularity level 3. This again, points to the importance of incorporating spatial correlation in customer acquisition models at the correct level of granularity.

Comparing the spatial model that includes all granularity levels with the spatial model at the optimal level proves the value of simultaneously including all granularity levels. Whereas in the first model all neighborhood effects needs 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 on 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 at a 0.0001 significance level. The spatial interdependences measured by these two spatial lag effects are already covered by other spatial variables.
Such a model is able to improve the AUC with an extra 0.87% compared to the best spatial model based on a single weight matrix which means a total improvement of 3.60% 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.

4.3 Sample size effect

In an autologistic model, spatial interdependence is incorporated based on a spatial lag effect that represents the purchasing behavior of neighboring customers. However, at finer granularity levels the number of observations within such matrix can become too small, affecting the reliability of the spatial influence. As a result, the sample size of the dataset can have an influence on the effect of these spatial parameters. In order to investigate this, smaller samples of the original dataset are generated. Table 4 gives an overview of the different sample sizes examined. Each sample is generated by randomly selecting a number of observations from the original dataset. In this way, 10 datasets are created for each sample size. Except for sample size "100%", for which only the original dataset is used.

Table 4. Overview of sample sizes

Sample size	Average number of observations	Average number of events
2%	62871	563.30
4%	125742	1094.90
6%	188613	1671.90
8%	251483	2211.00
10%	314354	2754.10

20%	628708	5550.50
40%	1257415	11075.00
60%	1886122	16612.60
80%	2514829	22102.70
100%	3143536	27657.00

Similar as done for the original dataset, each of the 90 newly created samples are split into a training (70%) and a validation sample (30%). On each of these training samples, a traditional model, 7 “single level” and an “all levels” autologistic model are estimated. Next, based on the validation sample, the predictive performances of these models are calculated in terms of AUC. The average predictive performance per sample size is presented in Fig. 2. First of all, this figure clearly illustrates the value of large datasets. In general, this figure indicates for all models that the larger the sample size is, the higher the predictive performance on the validation sample. However, this effect is even larger for the spatial models than for a traditional logistic regression model. This illustrates again the importance of collecting enough data to construct reliable spatial lag effects. Secondly, the larger the sample size, the more granularity level 3 emerges as optimal granularity level. When the sample size becomes smaller, the difference in predictive performance with spatial models based on coarser granularity levels becomes smaller. For sample size “2%” and “4%”, the spatial models on granularity level 1 even outperform the level 3 models. In other words, the optimal granularity level tends to move to a coarser level as a result of the smaller sample size.

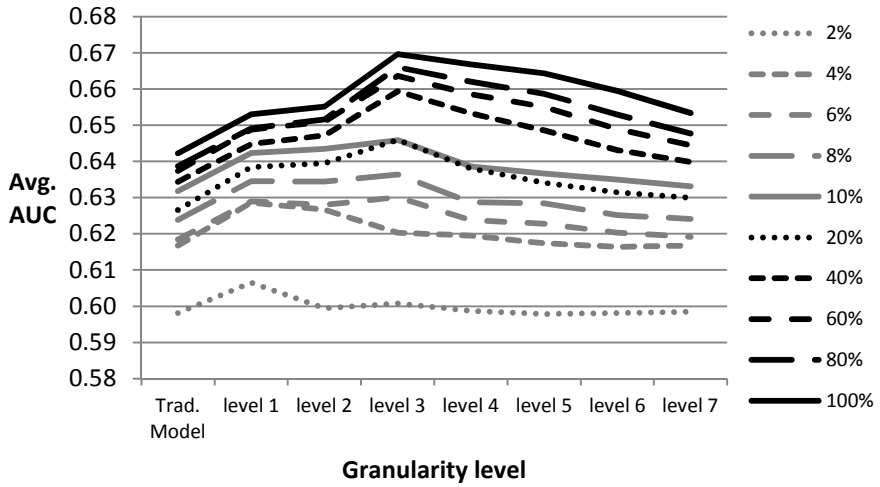


Fig. 2. Overview of the average AUCs at different sample sizes.

Fig. 3 explains this tendency by plotting the average spatial autoregressive coefficient of the "single level" autologisic models at several sample sizes. This figure shows that the sample size has an important effect on the spatial autoregressive coefficient (rho). In general, the spatial predictors become more important when the sample size increases. However, for the spatial autoregressive coefficient calculated on a coarse level of granularity, already a small data sample is sufficient to obtain a strong effect on the dependent variable. More specifically, for level 1 and level 2, the spatial autoregressive coefficient remains relative constant starting from sample size "6%". From this point, the spatial lag effects are constructed based on enough neighbors to be totally reliable. Similarly, the spatial autoregressive coefficient at level 3 flats out starting from sample size "60%". At this granularity level, more neighborhoods are used to incorporate spatial interdependence. As a result, more observations are needed to construct reliable spatial lag effects. The spatial variables constructed on even finer levels of granularity show a very small influence in the models based on small sample sizes, but once more data is available to construct better spatial lag effects, the impact of these spatial variables is clearly improving.

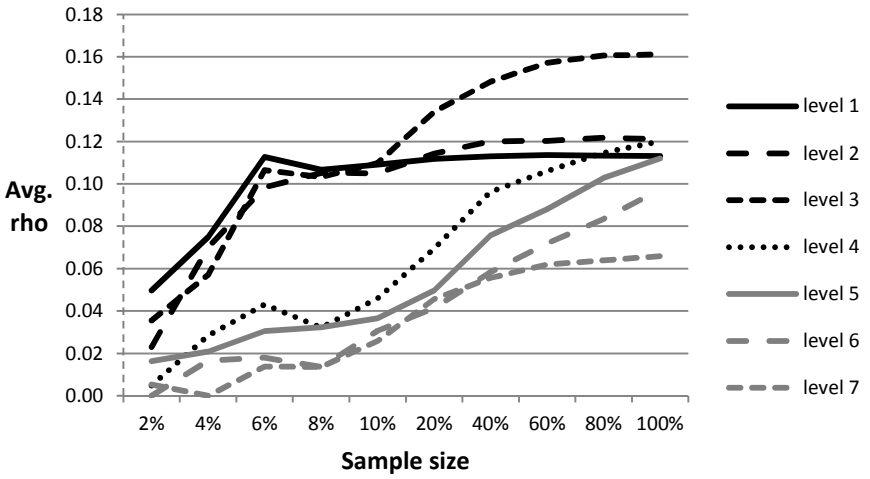


Fig. 3. Overview of the average spatial autoregressive coefficients (rho) of the "single level" autologistic models at different sample sizes.

Table 5. Comparison of the Average AUC between "single level" and "all levels" autologistic model at different sample sizes

Sample size	Avg. AUC best "single level" autologistic model	Avg. AUC "all levels" autologistic model	AUC difference
2%	0.6065*	0.6027	-0.0038
4%	0.6286*	0.6294	0.0008
6%	0.6301**	0.6357	0.0056
8%	0.6364**	0.6406	0.0042
10%	0.6458**	0.6508	0.0050
20%	0.6459**	0.6506	0.0047
40%	0.6594**	0.6644	0.0050
60%	0.6636**	0.6696	0.0060
80%	0.6660**	0.6730	0.0070
100%	0.6696**	0.6783	0.0087

* Based on level 1 model
** Based on level 3 model

Finally, the effect of the sample size is also examined for an autologistic model that simultaneously incorporates all levels of granularity. Table 5 makes a comparison of the predictive performance between such model and the best performing "single level" autologistic model at multiple sample sizes. Again the predictive performance is expressed in terms of the average AUC over 10 randomly created

datasets for each sample size. For sample size “100%”, only the original dataset is used. For sample size “2%” and “4%”, the level 1 model emerges as best performing “single level” model. Starting from sample size “6%” the data sample is large enough for the level 3 model to become superior. In the last column of Table 5 the difference between both a “single level” and “all levels” model is demonstrated. This clearly shows that the larger the data sample, the more one can benefit from the advantages of the extended autologistic model. At small sample sizes an “all levels” model is not able to outperform a “single level” model. At the smallest sample size these models perform even worse than a “single level” model. This is because on the training sample spatial variables created at finer granularity levels can become significant, but these variables have more the tendency to disturb predictions on the validation sample because they are not reliable enough. Once the data sample become larger, the predictive improvement, as a result of including multiple levels of granularity simultaneously, increases gradually.

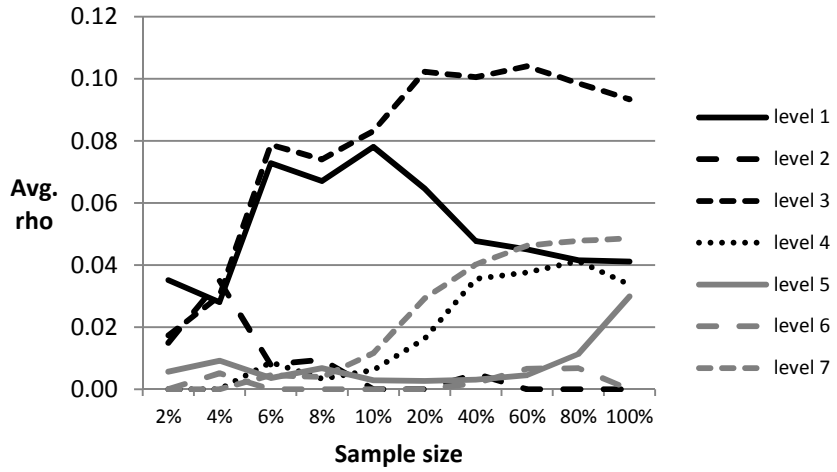


Fig. 4. Overview of the average spatial autoregressive coefficients (rho) of the “all levels” autologistic models at different sample sizes.

Fig 4. Explains this evolution by graphically representing the average spatial autoregressive coefficients of these extended autologistic models. This figure show a similar trend as observed in the “single level” models. The autoregressive coefficients at a coarser level

become quickly powerful at small sample sizes. When more data becomes available also the spatial variables calculated on a finer granularity level are gaining importance. By this, the model is better able to distinguish several origins of spatial interdependence using multiple spatial weight matrixes, resulting in an increasing improvement of predictive performance. Actually, this graph shows that once enough data is available to construct more reliable spatial lag effects at a finer granularity level, some of the spatial interdependence that is firstly explained by the level 1 spatial variable can be better explained on a finer level of granularity. In contrast to Fig. 3, some spatial autoregressive coefficients remain low in the “all levels” autologistic model because the spatial interdependences measured by these spatial variables are already covered by other spatial variables.

5 Discussion and conclusion

Traditional customer acquisition models often ignore the spatial correlation that could exist between the purchasing behaviors of neighboring customers and treats this as nuisance in the error term. Based on data of a Japanese automobile brand, this study shows that, even in a model that already includes a high number of socio-demographic and lifestyle variables typically attracted for customer acquisition, extra predictive value can still be obtained by taking spatial interdependence into account using a generalized linear autologistic regression model.

Further, 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 on a coarser granularity levels include too much interdependence that does not exist in reality, affecting the validity of the model. Though, if the level of granularity becomes too fine, the number of observations and events in each neighborhood declines, which can affect the reliability of the model. 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 proves that a model that simultaneously includes multiple granularity levels is able to outperform the best generalized linear autologistic regression model based on a single weight matrix. Hence, if the marketing decision maker has sufficient recourses it is advisable to obtain data which divides customers into neighborhoods at multiple granularity levels. This simplifies the decision to select optimal neighborhood level because this model is able to simultaneously incorporate all levels and automatically divide the existing interdependence, this causes each underling effect to be estimated based on its optimal granularity level.

In a sensitivity analysis, this study demonstrates how the sample size can influence the effect of the spatial variables. Spatial influences are included based on a spatial lag effect that incorporates the purchase behavior of surrounding customers living in the same neighborhood. Hence, this study shows that using a finer level of granularity is only valuable when enough data is available. If not, the spatial lag effect will be calculated based on too few observations, which affects the reliability of this variable. Consequently, when the data sample becomes smaller, the optimal level of granularity tends to move towards a coarser level. In addition, this also affects the use of a model that simultaneously takes multiple granularity levels into account. In order to take advantage of the fact that each origin of spatial interdependence can be measured on its optimal level, reliable spatial lag effects need to be constructed even on finer levels of granularity. As a result, the difference in predictive improvement between such extended model and a “single level” autologistic model increases gradually when the data sample becomes larger.

Although this study provides interesting insights, there are still some recommendations for future research. This study is executed on a specific CRM model for a specific product. It examines the incorporation of neighborhood effects in a customer identification model that predicts automobile preferences for a Japanese automobile brand. In order to generalize the conclusions in this study, future research should verify these findings in different contexts. First of all, this highly visible and luxury good is a perfect example on which social influences and spatial interdependence can be suspected. Further research could also investigate the effect of the chosen granularity level in a context of less visible or luxury goods. Secondly, data

augmentation is crucial in customer acquisition models because no transactional information is typically available, but incorporating spatial autocorrelation could also be valuable in other CRM disciplines, such as customer development or churn models. Finally, this study points out that the choice of neighborhood level can have an important influence on the model's accuracy. This study already examined the influence of sample size on the optimal granularity level, but further research could search for other elements that might have an influence on this optimal level.

Acknowledgement

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

References

1. Ngai, E. W. T., Xiu, L., Chau, D. C. K.: Application of data mining techniques in customer relationship management: A literature review and classification. Expert Syst. Appl. 36, 2592--2602 (2009)
2. Petrison, L. A., Blattberg, R. C., Wang, P. : Database marketing past, present and future. Journal of Direct Marketing 7, 27--43 (1993)
3. Ling, R., Yen, D. C. : Customer relationship management: An analysis framework and implementation strategies. Journal of Computer Information Systems 41, 82--97 (2001)
4. Kamakura, W., Mela, C. F., Ansari, A. , Bodapati, A. ,Fader, P., Iyengar, R., Naik, P., Neslin, S., Sun, B. , Verhoef, P. C. , Wedel M., Wilcox, R.: Choice models and customer relationship management. Mark. Lett. 16,279--291 (2005)
5. Baecke, P., Van den Poel, D.: Improving Purchasing Behavior Predictions by Data Augmentation with Situational Variables. Int. J. Inf. Technol. Decis. Mak. 9, 853--872 (2010)
6. Bradlow, E.T., Bronnenberg, B., Russell, G.J., Arora, N., Bell, D.R., Duvvuri, S.D., TerHofstede, F., Sismeiro, C., Thomadsen, R., Yang, S.: Spatial Models in Marketing. Mark. Lett. 16, 267--278 (2005)
7. Bronnenberg, B.J.: Spatial models in marketing research and practice. Appl. Stoch. Models. Bus. Ind. 21, 335--343 (2005)
8. Bronnenberg, B.J., Mahajan, V.: Unobserved Retailer Behavior in Multimarket Data: Joint Spatial Dependence in Market Shares and Promotional Variables. Mark. Sci. 20, 284--299 (2001)
9. Bell, D.R., Song, S.: Neighborhood effects and trail on the Internet: Evidence from online grocery retailing. QME-Quant. Mark. Econ. 5, 361--400 (2007)
10. Moon, S., Russel, G.J.: Predicting Product Purchase from Inferred Customer Similarity: An Autologistic Model Approach. Mark. Sci. 54, 71--82 (2008)
11. Grinblatt, M., Keloharju, M., Ikäheimo, S.: Social Influence and Consumption: Evidence from the Automobile Purchases of Neighbors. Rev. Econ. Stat. 90, 735--753 (2008)
12. Manchanda, P., Xie, Y., Youn, N. The Role of Targeted Communication and Contagion in Product Adoption. Mark. Sci. 27, 961--976 (2008)

13. Yang, S., Allenby, G.M.: Modeling Interdependent Customer Preferences. J. Mark. Res. 40, 282--294 (2003)
14. Steenburgh, T.J., Ainslie, A.: Massively Categorical Variables: Revealing the Information in Zip Codes, Mark. Sci. 22, 40--57 (2003)
15. Baecke, P., Van den Poel, D.: Including spatial interdependence in customer acquisition models: a cross-category comparison. Expert Syst. Appl. 39, 12105--12113 (2012)
16. Baecke, P., Van den Poel, D.: Data augmentation by predicting spending pleasure using commercially available external data. J. Intell. Inf. Syst. 36, 367--383 (2011)
17. Arndt, J.: Role of Product-Related Conversations in the Diffusion of a new Product. J. Mark. 4, 291--295 (1967)
18. Allsop, D.T., Bassett, B.R., Hoskins, J.A.: Word-of-mouth Research: Principles and Applications. J. Advert. Res. 47, 398--411 (2007)
19. Keller, E.: Unleashing the Power of Word of Mouth: Creating Brand Advocacy to Drive Growth. J. Advert. Res. 47, 448-452 (2007)
20. Carl, W.: What's all the Buzz about? Everyday Communication and the Relational Basis of Word-of-Mouth and Buzz Marketing Practices. Manag. Com. Q. 19, 601--634 (2006)
21. Bikhchandani, S., Hirshleifer, D., Welch, I.: A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. J. Polit. Econ. 100, 992--1026 (1992)
22. Chen, Y., Wang, Q.I., Xie, J.: Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning. J. Mark. Res. 48, 238--254 (2011)
23. Hartmann, W.R., Nair N., Manchanda P., Bothner M., Dodds P., Godes D., Hosanagar K., Tucker C: Modeling social interactions: identification, empirical methods and policy implications. Mark. Lett. 19, 287--304 (2008)
24. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in Social Networks. Annu. Rev. Sociol. 27, 415--444 (2001)
25. McCrea, R.: Explaining sociospatial patterns in South East Queensland, Australia: social homophily versus structural homophily. Environ. Plan. A 41, 2201--2214 (2009)
26. McCullagh, P., Nelder J.A.: Generalized linear models. Chapman & Hall, London (1989)
27. Hosmer, D.W., Lemeshow S.: Applied Logistic Regression. John Wiley & Sons, New York (2000)
28. Augustin, N. H., Mugglestone M. A., Buckland S.T.: An Autologistic Model for the Spatial Distribution of wildlife. J. Appl. Ecol. 33, 339--347 (1996)
29. Hoeting J.A., Leecaster M., Bowden D.: An Improved Model for Spatially Correlated Binary Responses. J. Agric. Biol. Environ. Stat. 5, 102--114 (2000)
30. He, F., Zhou, J., Zhu, H.: Autologistic Regression Model for the Distribution of Vegetation. J. Agric. Biol. Environ. Stat. 8, 205--222 (2003)
31. Besang, J.: Spatial Interaction and the Statistical Analysis of Lattice Systems. J. Roy. Statist. Soc. Ser. B (Methodological) 36, 192--236 (1974)
32. Besang, J.: Statistical Analysis of non-lattice data. The Statistican 24, 179--195 (1975)
33. Huffer, F.W., Wu, H.: Markov Chain Monte Carlo for Autologistic Regression Models with Application to the Distribution of Plant Species. Biometrics 54, 509--524 (1998)
34. Wu, H.,Huffer, F.W.: Modelling the distribution of plant species using the autologistic regression model. Environ. Ecol. Stat. 4, 49--64 (1997)
35. Kim, Y.S.: Toward a successful CRM: Variable selection, sampling, and ensemble. Decis. Support Syst.41, 542--553 (2006)
36. Hanley, J.H., McNeil B.J.: The meaning and use of area under a receiver operating characteristic (ROC) curve. Radiology 143, 29--36 (1982)
37. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or morecorrelated receiver operating characteristic curves: a nonparametric approach. Biometrics. 44, 837--845 (1988)