The impact of geographical factors on churn prediction: An application to an insurance company in Madrid's urban area

De la Llave, Miguel Ángel; López, Fernando A.; Angulo, Ana

Abstract:

Geography has previously been noted as a decisive factor in business literature. This paper provides evidence of the significant role geography plays in customer lapse behaviour in an urban environment. This novel approach is based on the idea that the customers who cancel all policies and leave the company are not randomly distributed; rather, a mimetic performance of close individuals is noted. The physical proximity of the customer to the geographical focus (strategical centre, as insurance offices) and the interaction with nearby customer are spatial factors that increase (or decrease) the probability of churning. An empirical analysis using more than 7,000 spatially georeferenced offline customers of a Spanish insurance company in the urban area of Madrid (Spain) demonstrated that the customer's proximity to offices of such insurance company under study decreases the probability of churning, whereas high lapse risk was detected in customers in the surroundings of the company's competitor branches. In addition, we identified spatial autocorrelation in churn probability, thus demonstrating that the probability of churn of a customer increases if nearby customers churn.

Keywords: lapse prediction, churn prediction, insurance, spatial autocorrelation, spatial probit model, Madrid

1. Introduction

The impact of geography on marketing science is an important topic of research for business and management. A model becomes '*spatial*' if the behaviour of one economic agent is codetermined by nearby economic agents (Burridge et al., 2016). Spatial analysis is a new and emerging research topic in marketing – one which has not yet revealed its potential – that is attracting more interest each day due to the increasing availability of georeferenced information. In the field of Customer Relationship Management (CRM), taking advantage of the spatial correlation between customers can improve the predictive performance of models. The main contributions to CRM (including spatial effects) are in the subfield of customer acquisition (Baecke and Van den Poel, 2012, 2013; Millo and Carmeci, 2011). However, in relation to customer churn behaviour no research that takes into account geography and 'space' as explicative factors has yet been undertaken.

Customer churn[1] prediction models aim to detect customers with a high propensity to leave. Because there are many competing companies, customer loyalty to a particular company has decreased, and high percentages of customers cancel all their policies. The percentage of churn ranges between 3.3% (Hung et al., 2006) and 15.7% (Keramati et al., 2014), and in other cases is confidential (Günther et al., 2014). Losing a customer has several negative effects on the company: first, the churning has implications for sales revenue. The cost of attracting new customers to replace those who have left is high. Some research has shown that this costs between 6 (Verbeke et al., 2011) and 12 times that of retaining the existing customer (Torkzadeh et al., 2006); Secondly, lost customers have a negative effect on the company's reputation and impact negatively on the brand's image. Churners tend to give negative feedback about the company, which may influence potential customers (Saradhi and Palshikar, 2011).

[1] 'Churn' is derived from two words, change and turn (Lazarov and Capota, 2007).

Therefore, predicting policy cancellation before the renewal date is a critical point for most companies. If such group of customers or policies can be detected, wherever the risk of churn is high, specific marketing actions (e.g. customer retention programmes) can be developed in order to retain the customers. A small decrease in retention rates should therefore provide the company with future benefits; it is clear that customer retention is a critical aspect of CRM.

The phenomenon of customer churn can be frequently observed in volatile consumer service markets such as telecommunications (Archaux et al., 2004; Hung et al., 2006; Rosset et al., 2003), insurance (Günther et al., 2014; Risselada et al., 2010; Morik and Köpcke, 2004), subscription services (Coussement and den Poel, 2008), financial services (Lariviere and den Poel, 2005) and banking (Xie et al., 2009). A huge variety of methodological approaches have been discussed in examinations of market independence. The most popular of these approaches uses classification trees (Lemmens and Croux, 2006) and logistic regression (Günther et al., 2014); multiple statistical techniques[2] have also been developed in order to identify customers who are likely to churn based on their characteristics: for instance, survival analysis (e.g. Brockett et al., 2008); neural networks (Hung et al., 2006); random forest (Lariviere and Van del Poel 2005); support vector machines (Xie et al., 2009); and more recently, machine learning (bagging; boosting; staking; voting) has been applied (Risselada et al., 2010). Most of those techniques have resulted in limited gains in accuracy and substantial increases in complexity (Risselada et al., 2010). This statement is also supported by Neslin et al. (2006), who found that logistic regression models and classification trees accounted for 68% of entries in churn modelling.

[2] A full description of methodologies used in the churn prediction model, besides the most important contributions, is to be found in Table 1 in Verbeke et al., (2011); Table 1 in Soeini and Rodpysh (2012); Table 1 in Abbasimehr et al., (2014); Table 1 in Allahyari and Vahidy, (2012); and Table 1 in Tsai and Lu (2009). A comparative study is presented in Vafeiadis et al. (2015).

Most insurance companies collect very large data sets that provide invaluable business information which may be analysed to develop a better understanding of customer behaviour. In some cases, the companies have several million customers, and they store a huge number of attributes for the holder of each policy underwritten, mainly socio-demographic characteristics (education level, age, sex, family size, social status) and specific information about the company's relation with the customer (number of policies, discount programme), and even relationships with another customers (social network, family relations between customers). There is one piece of information, which is included in all data warehouses that, to the best of our knowledge, has never previously been used in models of churn prediction: the address of the customer. The address of a customer is an important piece of information that enriches any churn model. First, if the insurance company knows the neighbourhood (or zip code) of the customer, then indirectly such company could collect information about the customer's economic status, allowing to divide customers into limited neighbourhoods. Some research undertook on zip codes in churn prediction (Löchl et al. 2009; Verbeke et al., 2012; Huigevoort and Dijkman, 2015) showed ambiguous conclusions. Verbeke et al. (2012) writes, "*the number of times a customer called the helpdesk will most probably be a better predictor of churn behavior than the zip code*". Secondly, and directly related to this research, knowing the exact location of a customer (latitude and longitude coordinates) makes it possible to identify the proximity of other customers. Nearby customer churn behaviour is probably codetermined, and some mimetic conduct between them can be detected. Pinheiro and Helfert (2010) wrote, "*Some events within the network can be influenced by activities of other customers. In the example of churn, word of mouth, rumors, commentaries and mostly activities of churn of other customers may create a chain process*". Along the same lines, Haenlein (2013) presents evidences on the importance of social interaction in customer churn decisions. Lastly, if the company knows the exact location of a customer, it is easy to identify geographical factors (strategic geographical points) that could be related to churning. Although there is a high degree of heterogeneity in insurance distribution channels, proximity to a tied-agent (or insurance office) of the company (or the competition) is probably a factor that influences churn (Christiansen et al., 2014). The predominant distribution channel for the larger insurance companies in the individual market is often still the tied-agent channel (Dumm and Hoyt, 2003, 28). Tied-agents are paid by a particular insurance company to sell only its products. The presence of the agent's company is the only variable in the model which is directly controllable by the company's managers; they might have only indirect control over the market share of the broker and the direct distribution channels (Löchl et al., 2009). In this sense, "*a systematic analysis of spatial information can identify profitable locations. Since the cost of analysis is relatively low, it would appear worthwhile for financial service firms to invest in a systematic analysis of locational and demographic factors*" (Clapp et al., 1990, 447).

On the other hand we recognise that ecommerce is an emergent channel and the percentage of customer buying insurances through online channels will increase in the future in all likelihood. The behaviour of online customer is different from the offline customer. That different behaviour is reported in the literature mainly with respect to the purchase behaviour (e.g. Mau et al., 2017) and lapse behaviour (e.g. Boehm, 2008; Christiansen et al., 2014). In this sense, some assumptions suggest that the competition on the Internet is only a few mouse clicks away and therefore online customers have more probability to lapse (Boehm, 2008). In contrast to this, the offline environment offers more opportunities for personalized marketing as well as greater flexibility and convenience to the customer and therefore the offline customers have less probability to lapse. Christiansen et al. (2016, 16) confirm this hypothesis '*We find that contracts purchased from tied agents are less likely to lapse*'. Finally, for the online customers it more difficult has a clear geographical localization.

Taking into account the state of research, the main objective of this paper is to demonstrate the impact of geographical factors on churn prediction. Using the information provided by a local Spanish company, we will prove the power of using geographical information to improve the classical probit regression models using Spatial Regression Probit Models (LeSage, 2009). We selected this methodology based on the Spatial Autoregressive Probit model since the usual probit model is a popular methodology that has shown to perform well in churn analyses. Moreover, the parameter estimations are easily interpretable. It is to be born in mind that geographical factors improve the performance of most aforementioned methods. This paper fills an important lacuna in the literature, and it will be a turning point in churning.

The paper is structured as follows: the second section describes the data and methodology; the third section presents the most important results and some potential companies' strategies and the last section concludes this work.

2. Data and methodology

2.1 Data

The information used in the analysis comes from a Spanish insurance company which offers a wide range of insurance lines. Data analysed in this paper refers to those customers who have taken out at least one policy through the company's branches. For this work, we selected only customers located in the municipality of Madrid (Spain) and focused on the year 2015. We selected Madrid because the company present the greatest insight into that insurance market and because the number of policies in that area is the highest of any urban environment. A total of 11626 customers were selected in this first filter. The addresses of all customers were obtained from original data, and the exact coordinates (latitude, longitude) were integrated in the database. A geo-referencing process was carried out using the R package ggmap. Observations not properly localised or poor geocoding addresses were excluded from the sample (1817 data excluded in this filter). Lastly, customers who contracted their policies via ecommerce (online customers) are excluded because their lapse behaviour is different (Boehm, 2008). As a result of the aforementioned filters, the final dataset consists of a sample of 7,302 offline customers (customers from now on). Additionally, the addresses of all insurance agencies of the target company and the more relevant insurance competing agencies were obtained. The coordinates of all agencies were obtained using similar procedure. Figure 1 shows the analysed urban area and the spatial distribution of cases.

--- Figure 1 around here ---

Churn occurs when a customer leaves the company, cancelling all his policies. The overall lapse for the whole sample portfolio is 11.8% which is, according to statistics published by the insurance institute ICEA[3], similar to figures for other companies in the Spanish insurance market. These cancellations mainly consist of non-payments or voluntary surrenders. Client's reasons for cancellation are unknown because there is no any further questionnaire asking for possible triggers.

In order to predict customer behaviour, we selected a set of explicative factors. We included factors that have been considered significant in similar studies (e.g. Günther et al. 2014; Risselada et al., 2010). These factors are mainly related to the sociodemographic characteristics of customers and the contractual terms of their policies. Also, using information about the exact location of customers and agencies, we

[3] Cooperative Research between Insurance Companies (ICEA) (2016) publishes figures for cancellations in this industry annually (http://www.icea.es/).

detected other geographical variables which we think that could be relevant to this research. The description of all the analysed variables can be found in Table 1.

As regards the socio-demographic characteristics of customers, our dataset collates information on gender, age and the customers' familial status. Almost 60% of the customers are male; the age of the client and the familial status is information which is gathered when the customer signs his/her policy with the company.

Regarding the contractual terms of customers, particular attention was paid to the duration of the customer-company linkage (Years), the type of policies that the customer holds with the company at the beginning of 2015 (Policies), as well as the sum of the premium of all of them (Premium). The average duration of the relationship between the customer and the company is 5.23 years. Most customers (80%) have only one policy with the company. The predominant types of insurance in the portfolio analysed are car (37%) and home (60%) basically because the first one is mandatory in Spain and the second one is highly recommended when a mortgage is established. The remaining policies belong to health, funeral and pet insurances, which are residuals or merely complementary to the others. The average premium paid by the customers for their different insurance policies is €487 per year.

Information provided by the insurance company is enriched by geographical information such as the customers' addresses. First, for each customer, we defined his/her distance to the nearest analysed company agency (Dist-Own) as well as his/her distance to the nearest competitor agency (Dist-Compet). The average distance to an analysed company agency was 964 meters whereas the average distance to a competitor agency was 628 meters. Additionally, we introduce a scoring variable of the level of income per residential area in Madrid. This variable is published on a yearly basis by public institutions in Madrid and goes from one to ten, where ten is the maximum income score. The average scoring for the areas analysed is 6.84.

--- Table 1 around here ---

2.2 Methodology

Discrete choice models are popular tools to explain the effects of various factors in observed choices. It is useful to begin with a brief discussion of general binary response models before the addition of any spatial dependence pattern. In this subsection, we present the methodology for classical probit and spatial probit models.

Let *Y* be a binary Nx1 vector that reflects information on whether or not customers have churned during a certain period; that is:

$$y_i = \begin{cases} 1 & \text{if customer i leaves company (customer i churns)} \\ 0 & \text{if customer i does not leave company (customer i does not churn)} \end{cases} \qquad (1)$$

Note the difference in unobservable profits associated with the 1-0 choice indicators: $\eta_{1i} - \eta_{0i}$ $(i = 1,...,N)$, where η_{1i}represents customer i's profit when leaving the company and η_{0i} represents the customer's profit when he/she stays with the company. Let us define such differences as the unobservable (latent) variable y_i^*, which it is related to the observed variable y_i, as follows:

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* = \eta_{1i} - \eta_{0i} > 0 \\ 0 & \text{if} \quad y_i^* = \eta_{1i} - \eta_{0i} \leq 0 \end{cases} \quad (i = 1,...,N) \qquad (2)$$

That is, if the difference $y_i^* = \eta_{1i} - \eta_{0i}$ is positive, the customer will leave the company ($y_i = 1$); if the difference is negative, he/she will stay with the company.

Although y_i^* is not observable, it will be assumed that it is determined by a set of explicative variables.

As a starting model, we defined a **non-spatial probit model**, which assumes a lineal relationship between the unobserved latent variable y_i^* and a set of (k-1) non-spatial explicative variables:

$$y_i^* = \gamma_1 + \gamma_2 x_{2i}^{ns} + \ldots + \gamma_k x_{ki}^{ns} + \varepsilon_i^{ns} = x_i^{ns\prime}\gamma + \varepsilon_i^{ns}\,; \text{ In matrix terms, } Y^* = X^{ns}\gamma + \varepsilon^{ns} \qquad (3)$$

where X^{ns} denotes the corresponding (Nxk) matrix of covariates. It is also assumed that the perturbance term ε^{ns} follows a standard normal distribution with a variance equal to 1 for identification purpose, $\varepsilon_i^{ns} \equiv N(0,1)\ \forall i$. The perturbance term is used to denote that two customers with the same characteristics can make different choices.

Among the so-called ‘non-spatial’ explicative variables, we consider as variables the socio-demographic characteristics of customers in Table 1 (Gender, Children and Age) together with the contractual term variables also shown in the table: Policies, Years and Premium.

As expressed in the Introduction, our hypothesis is that the geographical location of customers plays a relevant role in choice outcomes. If this is the case, the omission of such information in the model would lead to spatial dependence in the residuals of the estimated model and, even more importantly, the obtained estimated parameters would be inconsistent and inefficient (McMillen, 1992). The null of no spatial dependence in residuals of the non-spatial probit model in (2) can be tested by generalised Moran’s I statistic, as proposed by Kelejian and Prucha (2001) (see Amaral et al. 2013 for another alternatives). If the null of no spatial dependence were rejected by the data, alternative spatial probit specifications should be proposed.

2.2.1 Type I spatial probit model

Our first proposal is a spatial probit model which we denote as **type I spatial probit**, and which in fact extends the previous specification proposal. To be precise, we propose to extend model (3) using the geographical variables in Table 1: Dist-Own and Dist-Compet which represent the distances between a customer and an analysed company office and a competing one, respectively. The log of both variables will be included in the model to reduce the level of heterogeneity. The extended model can be expressed as follows:

$$y_i^* = x_i^{ns\prime}\gamma + x_i^{s\prime}\delta + \varepsilon_i \; = x_i{}'\beta + \varepsilon_i \;; \qquad \text{In matrix term, } \; Y^* = X\beta + \varepsilon \tag{4}$$

Model (3) is nested in the model expressed in (4). However, from an estimation and interpretation point of view, they do not present any difference. For the notation of model (4), customer *i* leaves the company with probability *P*$_i$, which can be expressed as follows:

$$P_i = P\left\{y_i = 1 \middle| x_i\right\} = \Phi\left(E\left(Y_i^*\right)\right) = \Phi\left(x_i'\beta\right) = \int_{-\infty}^{x_i'\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \tag{5}$$

where $\Phi(\cdot)$ refers to the cumulated distribution of the normal distribution; it introduces a non-linear relationship between changes in the expected probability of churning and changes in the explicative variables.

$$\frac{\partial P}{\partial x_f} = average\left(P\left\{y_i = 1 \middle| x_{fi} = 1\right\}\right) - average\left(P\left\{y_i = 1 \middle| x_{fi} = 0\right\}\right) \tag{6}$$

2.2.2 Type II spatial probit model

Next, for a **type II spatial probit** proposal, we will discuss the **Spatial Autoregressive (SAR) probit** model, proposed by LeSage and Pace (2009). Following the notation for model (4), it reads as follows:

$$Y^* = \rho WY^* + X\beta + \varepsilon \qquad \varepsilon \equiv N(0, I_N) \tag{7}$$

where the spatial lag of the latent dependent variable WY^* involves the NxN spatial weight matrix W. From several definitions for *W* proposed in the existing literature[4], the row-standardisation of the *m*-nearest neighbour *W* matrix was adopted here. As is well-known, using this approach, the *W* matrix contains elements of either *1/m* or 0. If customer *j* represents one of the *m*-nearest neighbours to customer *i*, the $(i,j)^{th}$ element of W contains the value *1/m*. Otherwise, a value of zero would be assigned to that W element. This results in the (Nx1) vector WY^* consisting of an average of the *m* neighbouring consumers' utility, and it creates a mechanism for modelling interdependence in consumer churn choices. In model (7), it should be observed that choices in one location are likely to be quite similar to choices made at nearby locations. That is, the model takes into account the possible spatial spillover among neighbouring consumer choices. The scalar parameter ρ measures the strength of dependence, with a value of 0 indicating independence. Clearly the first type of spatial probit model emerges when ρ=0 and therefore model (4) is nested in the model (5)

For estimation of this model several alternatives has been proposed to estimate the parameters of a spatial probit model. The Generalised Method of Moment (Pinkse and Slade, 1998); maximum Likelihood using the Expectation-Maximization algorithm (McMillen, 1992) or Bayesian Gibbs sample approach proposed by LeSage (2000). In this paper we use the procedure based on conditional maximum Likelihood recently

[4] Mainly based on contiguity, neighbourhood or distance.

developed by Martinetti and Geniaux (2017) because is very efficient and reliable since conditional estimators outperform the respective full-likelihood estimators.

3. Results and discussion

First, in this study a univariate approach has been used to explore any possible nonlinear relationships between the independent variables and the churn frequency in the sample. Figure 2 depicts the lapse rates for each of the continuous variables, which are split into segments. Two variables exhibited a general positive relationship with churn rate (Premium, Dist-Own) and four variables a general negative relationship (Age, Year, Dist-Compet, Income). However, there was a nonlinear tendency in several cases. For example, a positive relationship was noted between lapse rate and premium paid by the insured, mainly in the intermediate segments, and a constant or decreasing tendency appears in the last segments. For the age variable, a strong negative tendency was noted for younger consumers (approximately under 45), that changed to a constant tendency for the consumers classed as being of medium age and older. That is, the percentage of lapse rates decreased for older consumers. In the case of the Income of the client’s area, a clear pattern of convexity was found. Finally, a non-defined, but clearly non-lineal, pattern was observed for the Year variable. With respect to the two distance variables (in log), both results exhibit what is close to a linear pattern.

--- Figure 2 around here ---

Secondly, taking into account the results depicted in Figure 2, the first step in the specification process is to choose the better functional form to capture the observed non-linear effects. To achieve this objective, we selected the Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991) approach (using the library earth of R (Milborrow, 2011)) in order to select the best specification. The first column in Table 2 describes the results with the relevant split of the explicative variables for our modelling purpose.

--- Table 2 around here ---

Specification diagnostics for the estimated model are shown at the bottom of Table 2. First, the area under the ROC curve (AUC) indicates that the model correctly predicted 73.6% of the choices, and therefore the baseline model exhibited an acceptable level of predictive performance. In general, the results show that the selected socio-demographic and contractual policy term variables are important when examining consumer churn behaviour.

Regarding the sociodemographic variables, results are as follows. The gender variable is significant, demonstrating that the presence of women tends to be more stable in the company than that of men. This finding is in line with the common stereotype, based on the widely published idea, that males exhibit lower levels of loyalty than females (Melnyk et al., 2009). Analogous work can be found in Günther et al. (2014), in which loyalty to an insurance company was tested and similar results emerged. The results relating to age imply that older customers become more loyal to the company, which newly confirms the results showed in Günther et al. (2014). An optimal breakpoint in this variable is 46 years old, based on MARS methodology. At this point the slope becomes less steep in the regression model. This finding could be related to the fact that young people are more active on the Internet in terms of looking for cheaper alternatives, or it could be related to the fact that differences in economic status by age could change the customer’s aversion to changing his/her insurance company. However, when the client had children, although this variable exerted a negative impact on churn probability, the effect was not significant (at 5% level of significance).

Contractual term policy variables are also relevant when examining customer lapse choices. Firstly, customers who pay a higher premium up to €420 are more likely to cancel their policies. This is the inflexion point found using the MARS methodology. From that point, when customers spend more money on insurance, they tend to be less loyal although to a minor extent. In the same context, when the customer holds more than one policy in the company, the likelihood of that customer leaving the company is reduced. This can be explained by the fact that the time spent on searching for better conditions in other companies increases when customer owns different insurance types. Likewise, the type of insurance is also important when calculating the probability of churning. Results show that car insurance increases the probability of losing a client, probably because this is a line of business where there is much competitiveness in market, so a big effort has to be made to retain this car's client for years. On the contrary, home, health and funeral insurances seem to be more stable businesses, where clients are more reluctant to change. This behaviour is reasonable because of the grace periods and medical examinations that theses insurances frequently ask.

Finally, the number of years as a customer (Years) was also linked to a negative effect on the likelihood of leaving the company, although the effect is not significant – at the 5% level.

As previously stated, we hypothesise that geography could also play an important role in consumer churn decisions. If that were the case, parameters in baseline models would be biased owing to the omission of a relevant variable. The use of more detailed information would notably improve the explicative and predictive performance of the model. In order to check spatial dependence in residuals of the baseline model, we used the generalised I Moran test (Kelejian and Prucha, 2001). In order to use this test to check the hypothesis of spatial independence between customers preferences is necessary previously define the connectivity criterion using the W matrix. We select the criterion of k nearest neighbourhood (knn) connecting each observation with the k nearest (k=10, 15, 20) because the codeterminate behaviour has a local effect. The value of this test, displayed at the bottom of Table 2 for different connectivity criteria (W_{10nn}; W_{15nn}; W_{20nn}) indicates that spatial autocorrelation in the baseline model is noticeable. Therefore, this model should benefit from the use of spatial information. Consequently, we next propose the estimation of the so-called '**type I spatial probit model'**, and take into consideration a scoring for the income of the residential area of the client and the effect of geographical distance on the customer and insurance agencies (the analysed company's own and the competitors'). The results, which are in the second column of Table 2, are clear: firstly, individuals living in poorer zones tend to be less loyal and cancel their policies more than richest areas. This is evidence of the purchasing power of different areas of Madrid and the lifestyle of people living in. Secondly, the insurance company is more likely to lose a customer if it is located near a competence agency or it is far from an analysed company agency. Both results are reasonable, since the agent plays an important role in customer linkage. The net number of branches of an insurance company is essential to sell the product and to keep it in the portfolio. Also, we think that these results are related to the spatial positioning strategy of the company. Those areas where the company is not located but the competing agency is, are potential churning zones.

The relevance of the two geographical variables included in the type I spatial probit model can be deduced from the increase in the area under the ROC curve area. Also, socio-demographic and contractual variables in the type I spatial probit model are in accordance with those derived from the non-spatial probit model. However, the tests for spatial independence reject the null. As explained in the methodological section, our treatment of the autocorrelation problem in the data was based on the estimation of the **type II spatial probit model** or, more precisely, the proposed **SAR probit model**. As previously stated, in this model the aim is to take into account that the decision of a customer can be affected by the decision made by another nearby customer. The importance of such effects is captured through the new estimated parameter (ρ in Table 2), which is denoted “spatial autoregressive coefficient”. Results for the proposed SAR probit model, considering the effects of the customer’s k=10 nearest neighbours, are detailed in the last column of Table 2. We use the ML methodology propose by Martinetti and Genoaux (2017). The estimated spatial autoregressive coefficient is positive and significant. Similar results have been obtained for different numbers of neighbours proposed. This result confirms the existence of a positive and significance contagious effect, or spillover effect. The significance of the new ρ parameter also means that Model 3 outperforms the previous nested Model 2. The better performance of our final SAR probit model is reflected in a higher value of the area under the ROC curve. Figure 3 shows the ROC curves of three models. On the left side the ROC curve of Type II spatial model exceeds in almost all points the ROC of Baseline model and on the right side the ROC curve of Type II spatial model is similar to ROC curve of Type I spatial model but slight high. Although it seems to be a small increase, this improvement could represent large economic revenue for the company since, with the correct predictions, marketing managers can avoid the loss of some of the company’s customers. The stability of the effects of sociodemographic and contractual variables denotes a high level of robustness in our results.

3.1 Validation of model

Some information regarding the performance of the models under an out of sample validation exercise is presented in this subsection. Cross-validation exercises in a spatial framework present some problems (Le Rest et al., 2014). Note that a critical requirement is that the training and validation subsamples must be independent. In a spatial context all observations are related to each other and therefore the usual procedure of selecting random subsets to split the sample is not satisfactory. In order to minimize this problem in our cross-validation exercise we select compact training and validation subsamples. By selecting one location randomly across the map we split the database into two parts. On the one hand, the test subsample, which consists of 40% of the data, is formed by choosing the nearest neighbours of the selected location and on the other hand the rest 60% of the data which is used to train the three models. This process is repeated for one thousand times. Therefore, by following this methodology we ensure that as far as possible we do not break the embedded spatial dependency between the observations, so training and test samples are spatially independent to each other.

The main purpose of this subsection is to obtain the variability of the ρ parameter estimated in model 3 and the improvement in AUC indicator measured in the test subset. In Figure 4.a the boxplot of ρ estimation is depicted where all ρ values are positive and significant, confirming the importance of spatial effects in the training subsample. Figure 4.b shows the boxplots of the AUC estimation for 1,000 test subsamples. The results show that the AUC for Type II spatial probit model is slightly high that the Baseline model and also model 2. (AUC Baseline Model=0.7165, AUC Type I Model=0.7176, AUC Type II Model=0.7220).

Finally, from the selected SAR probit model we can draw important conclusions not only on the effect of changes of a variable on a respective customer (direct effect) but also on the rest of the customers (indirect effect).

3.2 Interpreting effects in a spatial probit model

Interpreting the way in which changes in the explanatory variables impact on the probability of churn is easy for the classical probit models while requires more care in the case of the SAR probit model expressed in (7). The reason is that, because of the spatial lag of the latent dependent variable WY^* of the SAR probit model, changes in the value of the variable for customer *j*, x_{hj}, influence customer *i*'s decision. That is, now, the changes to the probability of the churn of consumer *i* are twofold: i) that induced by a change in the own-value of the variable, $\frac{\partial P_i}{\partial x_{hi}}$,which is denoted in literature as the *direct effect*; and ii) that induced by a change in the value of the variable associated with another consumer, $\frac{\partial P_i}{\partial x_{hj}}$, denoted as *indirect effect*. Finally, a global effect measure, denoted *total effect*, gathers the sum of the direct and all indirect effects associated with all consumers who are not consumer *i*. The *total effect* in the SAR spatial probit model is comparable with the only effect derived from any standard probit model (and also the only effect derived from our first type spatial probit model). In essence, the idea is that spatial dependence expands the information set to include information on neighbouring individuals. A full description of interpretation of direct, indirect and total effects can be found in Lacombe and LeSage (2018).

Following this methodology, Table 3 illustrates the direct and indirect effects of the spatial probit model. First, it is important to highlight that the most relevant variable when determining churning is the number of policies held by the insured. There is a 16.8% fewer chance of losing a customer if he/she has more than one policy. Cross-selling is vital for companies, as it means good benefits and also because it makes customers more loyal.

Secondly, important information that the company should know about its customers (e.g. age, gender and familial status) should be noted. These factors have a meaningful impact on the probability of the customer leaving the company. For instance, every year up to 46 years of age means extra premium retention (increasing at a rate of 0.58% per year). This is clearly information which should be used to modulate and optimise the premium renewal every year depending on the personal features of the customer. Premium is also a variable in the process of optimisation. In our sample, we noted that up to €420, every €100 euros paid by the customer increases the likelihood of lapses by 2.1%.

--- Table 3 around here ---

Finally, in our research, we noted that the geographic position of the company plays a key role in the sustainability of an insurance portfolio. As can be ascertained using the information in Table 3, it is worth using geographic variables in the analysis, as it gives a sense of how dominant branches are with regard to the competition and how the company should treat its different clients depending on the area they have their residences. Short distances between agents and customers are crucial to maintaining long relationships with customers. In addition, in places where the company is not present, there is a potential risk for disengagement amongst the customers. In our sample, an additional log (kilometre) of distance among customers and tied-agent increases the probability of churning by about 0.7%. This likelihood might be increased if competence has closer branches. An additional log (kilometre) of distance between customers and the competence reduces the churn probability to 1.0%.

4. Conclusions and business management implications

In order to better manage customer churn, companies need to fully understand the effect of the main determinants of churn customer choice. Although this important topic has been the focus of some attention previously in the literature, we think that recent methodological improvements, in relation to spatial econometric techniques, can help us to gain a better understanding of the problem.

Conventional econometric models for choices assume independence among consumer decisions. This assumption could generate inexact estimations of parameters that might have an economic impact on the results. In an urban environment, it seems unrealistic that the individual choice of a customer to churn is not influenced by the decision of his neighbours. Those spatial spillovers could be explained by direct interaction between neighbouring customers or by the omission of relevant factors (with spatial structure in the model that could exhibit spatial dependence; Lesage and Pace, 2009).

Technological advances in geographic information systems (GIS) make collecting spatial data easier than ever before. Consequently, the possibility of spatial correlation among observations can be explored in order to achieve a better specification for a churn model. This was the case in this present paper; by paying attention to geographical information related to the addresses of the customers of a large insurance company in Madrid, we have reached a final spatial churn model that outperforms the non-spatial one in terms of both explicative and prediction power. Our results provide evidence that the probability of customer churn significantly increases if nearby customers churn, due to the spillover effect. Furthermore, the use of georeferenced insurance agencies has provided interesting conclusions regarding the effect of the closeness of tied-agents. On the one hand, an additional log (kilometre) of distance between customers and a company tied-agent increases the probability of churning by about 0.7%. An additional log (kilometre) of distance between customers and the competence reduces the churn probability by 1.0%. Hence, spatial distribution of consumers and agencies can be a cause of great concern for insurance managers.

As far as we know, the present paper is novel in that it pays attention to the non-linearity effect of socio-demographic and contractual policy term variables in the model. In accordance to the literature, our results indicate that, to cope with stable portfolios, tied-agents of the company should focus on younger male consumers who have contracted with the company more expensive and/or a higher number of insurance policies. Furthermore, the MARS methodology used in this paper reveals relevant additional information not discussed previously in the literature. The age of 46 represents an important breakpoint, since consumers below that age are more likely to leave the company by cancelling all insurance policies. In this paper, we made an important breakthrough in relation to the premium paid by consumers and to the number of contracted policies. Regarding premium effects, the results have shown that, up to a premium of €420, every €100 paid by the customer increases the likelihood of lapses by 2.1%.

Finally, three points relating to this paper and future approaches should be noted. First, our research focused on customer behaviour in one specific year. Further investigation is needed to introduce time level to the regression. Dynamism in people's conduct through time is not reflected in the analysis. Secondly, increasing the number of areas studied could lead to other interesting findings. In order to do this, a great deal of geo-referencing work on business premises needs to be undertaken. Lastly, in this paper we have demonstrated the importance of distance (in metres) between customers and agents when predicting churning. It might be a good approach to introduce time references to measure the distance in hours from point to point on the map. There is scarce literature on this topic and more research is required.

References

Abbasimehr, H., M. Setak, and M. Tarokh, 2014, A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction, *The International Arab Journal of Information Technology*, 11 (6): 599-606.

Allahyari, R. and K. Vahidy, 2012, Applying Data Mining to Insurance Customer Churn Management, *IPCSIT*, 30: 82-92.

Amaral, P.V., L. Anselin, and D. Arribas-Bel, 2013, Testing for spatial error dependence in probit models, *Letters in Spatial and Resource Sciences*, 6(2): 91-101.

Archaux, C., A. Martin, and A. Khenchaf, 2004, An SVM based churn detector in prepaid mobile telephony, in: *International conference on information and communication technologies: From theory to applications*. Proceedings. 2004 International Conference on Information & Communication Technologies.

Baecke, P., and D. Van den Poel, 2012, Including spatial interdependence in customer acquisition models: a cross-category comparison, *Expert Systems with Applications*, 39 (15): 12105-12113.

Baecke, P., and D. Van den Poel, 2013, Improving customer acquisition models by incorporating spatial autocorrelation at different levels of granularity, *Journal of Intelligent Information Systems*, 41 (1): 73-90.

Boehm, M., 2008, Determining the impact of internet channel use on a customer's lifetime. *Journal of Interactive Marketing*, 22(3), 2-22.

Brockett, P.L., L.L. Golden, M. Guillen, J.P. Nielsen, J. Parner, and A.M. Perez-Marin, 2008, Survival analysis of a household portfolio of insurance policies: how much time do you have to stop total customer defection?, *The Journal of Risk and Insurance*, 75 (3): 713-737.

Brun C., A.R. Cook, J.S. Huay Lee, S. A. Wich, L. Pin Koh, and L.R. Carrasco, 2015, Analysis of deforestation and protected area effectiveness in Indonesia: a comparison of Bayesian spatial models, *Global Environmental Change*, 31: 285-295.

Burridge, P., J.P. Elhorst, and K. Zigova, 2016, Group Interaction in Research and the Use of General Nesting Spatial Models (p. 223-258), in: J. P. LeSage, K. Pace, and B. Baltagi, eds, *Advances in Econometrics: Qualitative and Limited Dependent Variables*, 37 (Amsterdam: Elsevier).

Christiansen, M.C., Eling, M., Schmidt, J.P., and Zirkelbach, L., 2016, Who is changing health insurance coverage? Empirical evidence on policyholder dynamics. *Journal of Risk and Insurance*, *83*(2), 269-300.

Clapp, J. M., J. A. Fields, and C. Ghosh, 1990, An examination of profitability in spatial markets: The case of life insurance agency locations, *Journal of Risk and Insurance*, 57(3): 431-454.

Collingham, Y.C., R.A. Wadsworth, B. Huntley, and P.E. Hulme, 2000, Predicting the spatial distribution of non-indigenous riparian weeds: issues of spatial scale and extent, *Journal of Applied Ecology*, 37: 13-27.

Cooperative Research between Insurance Companies, 2016, Caída en el ramo de vida (number 1440). http://www.icea.es/.

Cooperative Research between Insurance Companies, 2016, Los Seguros Multirriesgo a diciembre. (number 1449). http://www.icea.es/.

Coussement, K., and D. Van den Poel, 2008, Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, *Expert systems with applications*, 34 (1): 313-327.

Dumm, R.E., and R.E. Hoyt, 2003, Insurance distribution channels: markets in transition, *Journal of Insurance Regulation*, 22 (1): 27-47.

Friedman, J. H., 1991, Multivariate adaptive regression splines, *The Annals of Statistics*, 19 (1): 1-67.

Günther, C.C., I.F. Tvete, K. Aas, G.I. Sandnes, and Ø. Borgan, 2014, Modelling and predicting customer churn from an insurance company, *Scandinavian Actuarial Journal*, 2014 (1): 58-71.

Haenlein, M., 2013, Social interactions in customer churn decisions: The impact of relationship directionality, *International Journal of Research in Marketing*, 30 (3): 236-248.

Heagerty, P.J., and S.R. Lele, 1998, A composite likelihood approach to binary spatial data, *Journal of the American Statistic Association*, 93: 1099-1111.

Huigevoort, C., and R. Dijkman, 2015, *Customer churn prediction for an insurance company*. Ph.D. Thesis

Hung, S.Y., D.C. Yen, and H.Y. Wang, 2006, Applying data mining to telecom churn management, *Expert Systems with Applications*, 31(3): 515-524.

Kelejian, H.H., and I. R. Prucha, 2001, On the asymptotic distribution of the Moran I test statistic with applications, *Journal of Econometrics*, 104(2): 219-257.

Keramati, A., R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi, 2014, Improved churn prediction in telecommunication industry using data mining techniques, *Applied Soft Computing*, 24: 994-1012.

Lacombe, D., and J.P. LeSage, 2018, Use and interpretation of spatial autoregressive probit models, *Annals Regional Science,* 60(1), 1-24

Larivière, B., and D. Van den Poel, 2005, Predicting customer retention and profitability by using random forests and regression forests techniques, *Expert Systems with Applications*, 29(2): 472-484.

Lazarov, V., and M. Capota, 2007, Churn prediction, *Bus. Anal. Course. TUM Computer Science*.

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., and Bretagnolle, V., 2014, Spatial leave ‐ one ‐ out cross ‐ validation for variable selection in the presence of spatial autocorrelation. *Global ecology and biogeography*, 23(7), 811-820.

Lemmens, A., and Croux, C., 2006, Bagging and boosting classification trees to predict churn, *Journal of Marketing Research*, 43 (2): 276-286.

LeSage, J.P., 2000, Bayesian Estimation of Limited Dependent Variable Spatial Autoregressive Models, *Geographical Analysis*, 32 (1): 19-35.

LeSage, J.P., R. K. Pace, N. Lam, R. Campanella, and X. Liu, 2011, New Orleans Business Recovery in the Aftermath of Hurrican Katrina, *Journal of the Royal Statistical Society A*, 174: 1007-1027.

LeSage, J.P., and Pace, R.K., 2009, *Introduction to Spatial Econometrics* (CRC Press)

Löchl, M., H.R. Hauri, and K. Axhaussen, 2009, *Agents, space and market shares: a Spatial analysis of the Swiss Insurance Market*. ETH, Eidgenössische Technische Hochschule Zürich, IVT, Institut für Verkehrsplanung und Transportsysteme.

Martinetti, D., and G. Geniaux, 2017, Approximate likelihood estimation of spatial probit models, *Regional Science and Urban Economics*, 64: 30-45.

McMillen, D.P., 1992, Probit with spatial autocorrelation, *Journal of Regional Science*, 32 (3): 335-348.

Melnyk, V., S.M. Van Osselaer, and T.H. Bijmolt, 2009, Are women more loyal customers than men? Gender differences in loyalty to firms and individual service providers, *Journal of Marketing*, 73 (4): 82-96.

Milborrow, S., 2011, Derived from mda: mars by T. Hastie and R. Tibshirani. Earth: Multivariate Adaptive Regression Splines, 2011. R package.

Millo, G., and G. Carmeci, 2011, Non-life insurance consumption in Italy: a sub-regional panel data analysis, *Journal of Geographical Systems*, 13 (3): 273-298.

Morik, K., and H. Köpcke, 2004, Analysing customer churn in insurance data–a case study, in: *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 325-336) (Heidelberg, Berlin: Springer).

Neslin, S.A., S. Gupta, W. Kamakura, J. Lu, and C.H. Mason, 2006, Defection detection: Measuring and understanding the predictive accuracy of customer churn models, *Journal of Marketing Research*, 43(2): 204-211.

Pace, R.K., and J. P. LeSage, 2017, Fast Simulated Maximum Likelihood Estimation of the Spatial Probit Model Capable of Handling Large Samples, in: *Spatial Econometrics: Qualitative and Limited Dependent Variables*. Published online: 01 Dec 2016; 3-34.

Pinheiro, C., and M. Helfert, 2010, Neural Network and Social Network to enhance the customer loyalty process, in: *Innovations and Advances in Computer Sciences and Engineering* (pp. 91-96) (Netherlands: Springer).

Pinkse, J. and M.E. Slade, 1998, Contracting in Space: An Application of Spatial Statistics to Discrete-Choice Models, *Journal of Econometrics*, 85: 125-154.

Risselada, H., P. C. Verhoef, and T.H. Bijmolt, 2010, Staying power of churn prediction models, *Journal of Interactive Marketing*, 24(3): 198-208.

Rosset, S., E. Neumann, U. Eick, and N. Vatnik, 2003, Customer lifetime value models for decision support, *Data Mining and Knowledge Discovery*, 7: 331-339.

Saradhi, V.V., and G.K. Palshikar, 2011, Employee churn prediction, *Expert Systems with Applications*, 38(3): 1999-2006.

Soeini, R., and K. Rodpysh, 2012, Applying Data Mining to Insurance Customer Churn Management, *International Proceedings of Computer Science*, 30: 82-92.

Torkzadeh, G., J. C. J. Chang, and G. W. Hansen, 2006, Identifying issues in customer relationship management at Merck-Medco, *Decision Support Systems*, 42(2): 1116-1130.

Tsai, C.F., and Y.H. Lu, 2009, Customer churn prediction by hybrid neural networks, *Expert Systems with Applications*, 36: 12547-12553.

Vafeiadis, T., K.I. Diamantaras, G. Sarigiannidis, and K.C. Chatzisavvas, 2015, A comparison of machine learning techniques for customer churn prediction, *Simulation Modelling Practice and Theory*, 55: 1-9.

Verbeke, W., K. Dejaeger, D. Martens, J. Hur, and B. Baesens, 2012, New insights into churn prediction in the telecommunication sector: A profit driven data mining approach, *European Journal of Operational Research*, 218(1): 211-229.

Verbeke, W., D. Martens, C. Mues, and B. Baesens, 2011, Building comprehensible customer churn prediction models with advanced rule induction techniques, *Expert Systems with Applications*, 38(3): 2354-2364.

Xie, Y., X. Li, E.W.T. Ngai, and W. Ying, 2009, Customer churn prediction using improved balanced random forests, *Expert Systems with Applications*, 36(3): 5445-5449.

Table 1: Description of the variables and statistics.

	Definition	Mean (std)	Range
	Dependent variable		
Churn	1 if the customer cancelled all policies (in 2015) = 0 otherwise	0.12 (0.32)	0/1
	Independent variables		
Socio-Demographic			
Gender	Gender of customer (1=female; 0=male)	0.60 (0.49)	0/1
Children	= 1 if the customer has children	0.25 (0.43)	0/1
Age	Age of the customer	49.7 (12.2)	[17-75]
Contractual terms			
Multi-Policy	= 1 if the customer has more than 1 policy	0.19 (0.39)	0/1
Years	Number of years as customer of the company	5.23 (2.73)	[1-10]
Premium	Total premium (in thousands €)	0.49 (0.52)	0/1
Car	= 1 if client has a car insurance policy	0.37 (0.48)	0/1
Home	= 1 if client has a home insurance policy	0.60 (0.49)	0/1
Health	= 1 if client has a health insurance policy	0.07 (0.25)	0/1
Funeral	= 1 if client has a funeral insurance policy	0.14 (0.35)	0/1
Pet	= 1 if client has a pet insurance policy	0.03 (0.18)	0/1
Geographical Variables			
Income	Average income per resident area.	6.84 (2.76)	[1-10]
Dist-Own	Distance in metres to the nearest analysed company office	964 (732)	[0-7351]
Dist-Compet	Distance in metres to the nearest office of competition insurance company	628 (444)	[0-7397]

Table 2. Probit and spatial probit: churn prediction $^{(a), (b)}$

	Non-spatial probit	Spatial probit	
	Baseline model	Type I spatial probit	Type II spatial probit (SAR probit model)
	Coeff *(z-value)*	*Coeff* *(z-value)*	*Coeff* *(z-value)*
Intercept	-0.865*** (-4.5)	-0.969*** (-5.0)	-0.727*** (-3.7)
Socio-Demographic variables			
Gender	-0.087** (-2.0)	-0.083* (-1.9)	-0.079* (-1.8)
Children	-0.039 (-0.8)	-0.041 (-0.8)	-0.037 (-0.7)
h(46-Age)	0.027*** (6.4)	0.027*** (6.4)	0.027*** (6.4)
h(Age-46)	-0.002 (-0.8)	-0.002 (-0.6)	-0.002 (-0.7)
Contractual terms variables			
Multi-Policy	-0.774*** (-3.9)	-0.760*** (-3.8)	-0.779*** (-3.9)
Years	0.001 (0.1)	0.000 (0.1)	0.000 (0.1)
h(0.420-Premium)	-0.982*** (-4.9)	-0.973*** (-4.8)	-0.974*** (-4.8)
h(Premium-0.420)	-0.060 (1.1)	0.066 (1.2)	0.065 (1.2)
Car	0.288* (1.7)	0.270 (1.6)	0.278 (1.6)
Home	-0.335** (-2.0)	-0.344** (-2.1)	-0.334** (-2.0)
Health	-0.409** (-2.2)	-0.414** (-2.3)	-0.413** (-2.3)
Funeral	-0.052 (-0.3)	-0.063 (-0.4)	-0.061 (-0.4)
Pet	0.017 (0.1)	0.009 (0.0)	0.008 (0.0)
Geographical Variables			
h(7-Income)		0.039*** (2.9)	0.033*** (2.7)
h(Income-7)		0.030 (1.4)	0.025 (1.3)
Log(Dist-Own)		0.044* (1.9)	0.033* (1.8)
Log(Dist-Compet)		-0.056* (-2.0)	-0.047* (-2.0)
ρ			0.186*** (-9.0)
Diagnostic test of Spatial dependence			
I Moran$^{(c)}$ (W_{10nn})	2.29**	1.84*	1.66
I Moran (W_{15nn})	1.87*	1.41	1.22
I Moran (W_{20nn})	1.89*	1.41	1.22
Diagnostic tests			
AIC	4752.4	4744.6	4733.5
LogLik	-2362.2	-2354.2	-2349.7
LR test		15.9***	9.0***
AUC	0.736	0.739	0.740

$^{(a)}$ h(c-X)=max{0,c-X}; h(X-c)=max{0,X-c], where X is the variable under analysis and c is the breakpoint detected using the MARS methodology.

$^{(b)}$ * indicates significance at 10%; ** indicates significance at 5%; and *** indicates significance at 1%.

$^{(c)}$ Generalised I Moran test.

Table 3. Direct, indirect and total effect of the spatial probit model

	Direct effect	Indirect effect	Total effect
	Socio-Demographic variables		
Gender	-0.0139	-0.0031	-0.0170
Children	-0.0064	-0.0014	-0.0079
h(46-Age)	0.0047	0.0011	0.0058
h(Age-46)	-0.0004	-0.0001	-0.0004
	Contractual terms variables		
Multi-Policy	-0.1371	-0.0309	-0.1679
Years	0.0001	0.0000	0.0001
h(0.420-Premium)	-0.1714	-0.0386	-0.2099
h(Premium-0.420)	0.0115	0.0026	0.0141
Car	0.0489	0.0110	0.0599
Home	-0.0588	-0.0132	-0.0720
Health	-0.0726	-0.0163	-0.0889
Funeral	-0.0108	-0.0024	-0.0132
Pet	0.0014	0.0003	0.0017
	Geographical Variables		
h(7-Income)	0.0058	0.0013	0.0071
h(Income-7)	0.0044	0.0010	0.0054
Log(Dist-Own)	0.0059	0.0013	0.0072
Log(Dist-Compet)	-0.0083	-0.0019	-0.0102

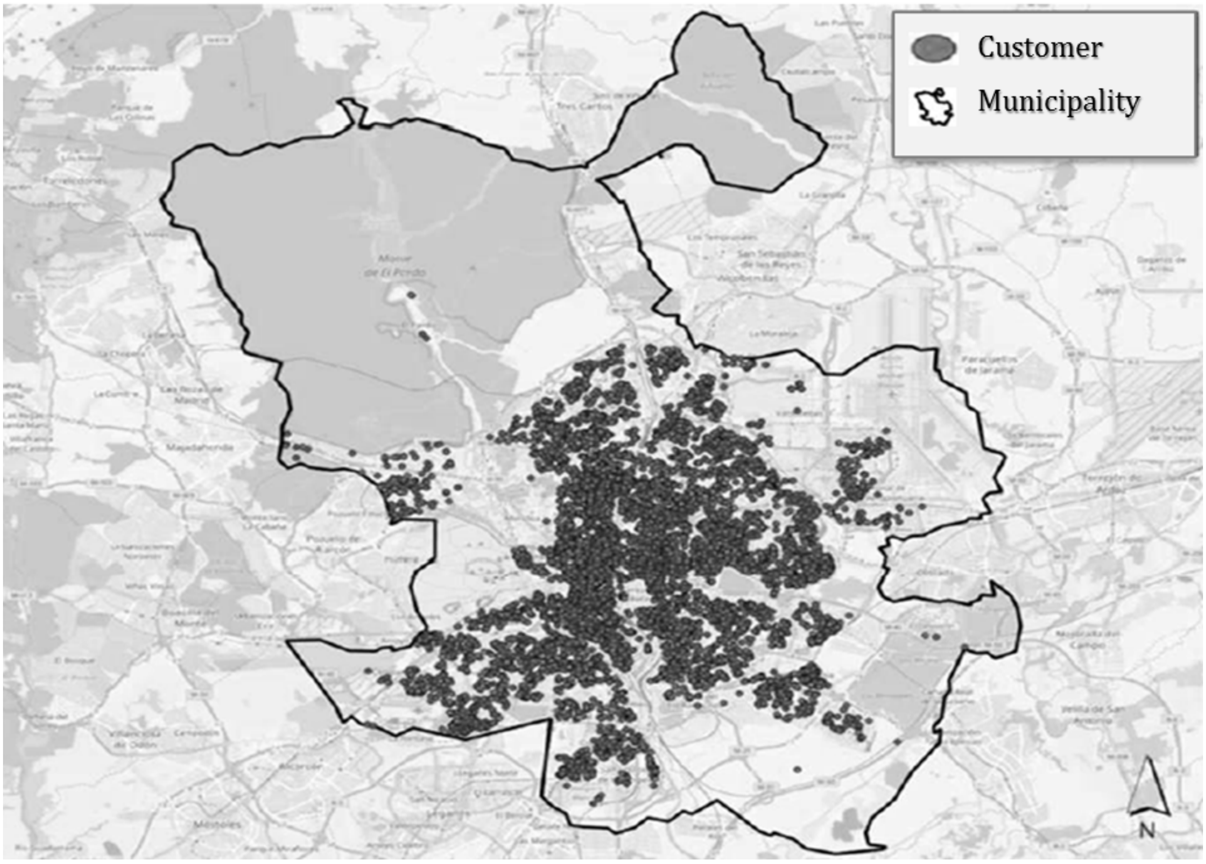

Figure 1a. Georeferenced customers

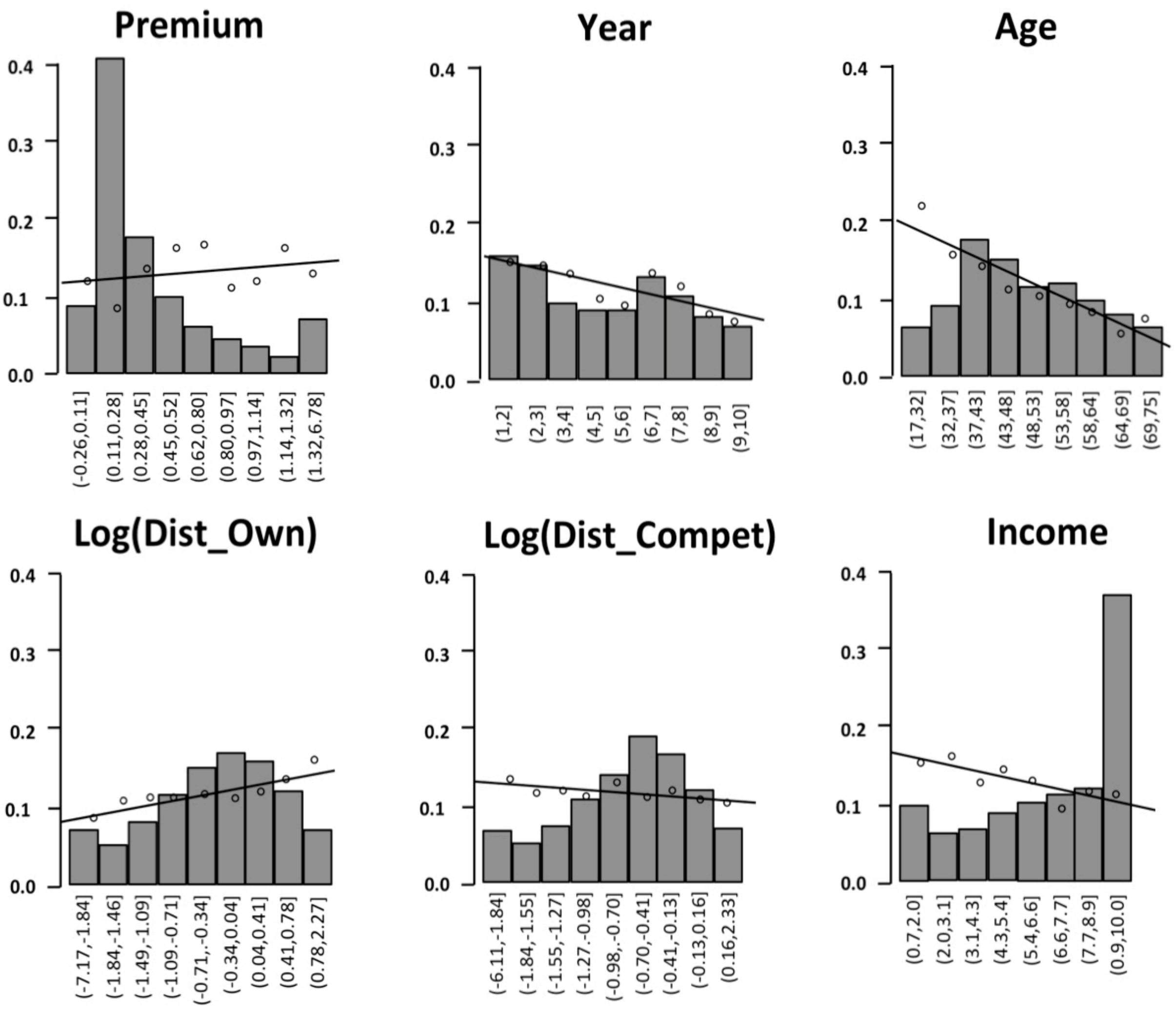

Figure 2. Lapse rates and histograms (in segments) for the continuous explicative variables.

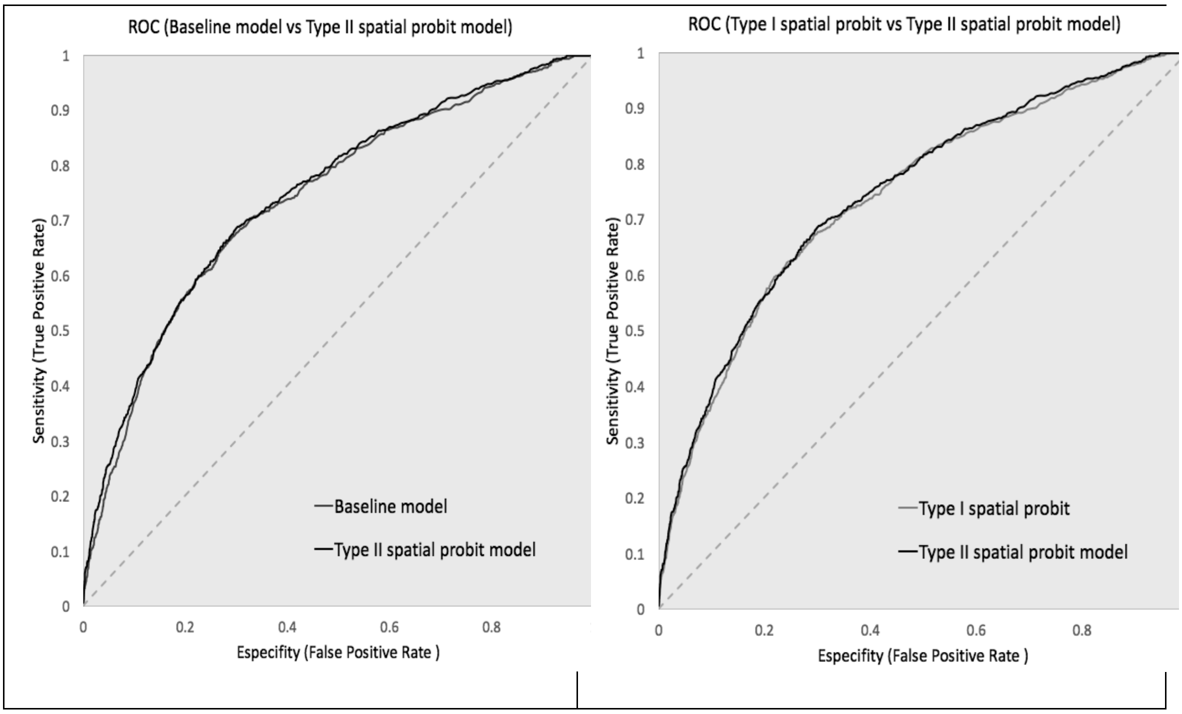

Figure 3. ROC curves comparing the three models estimated.

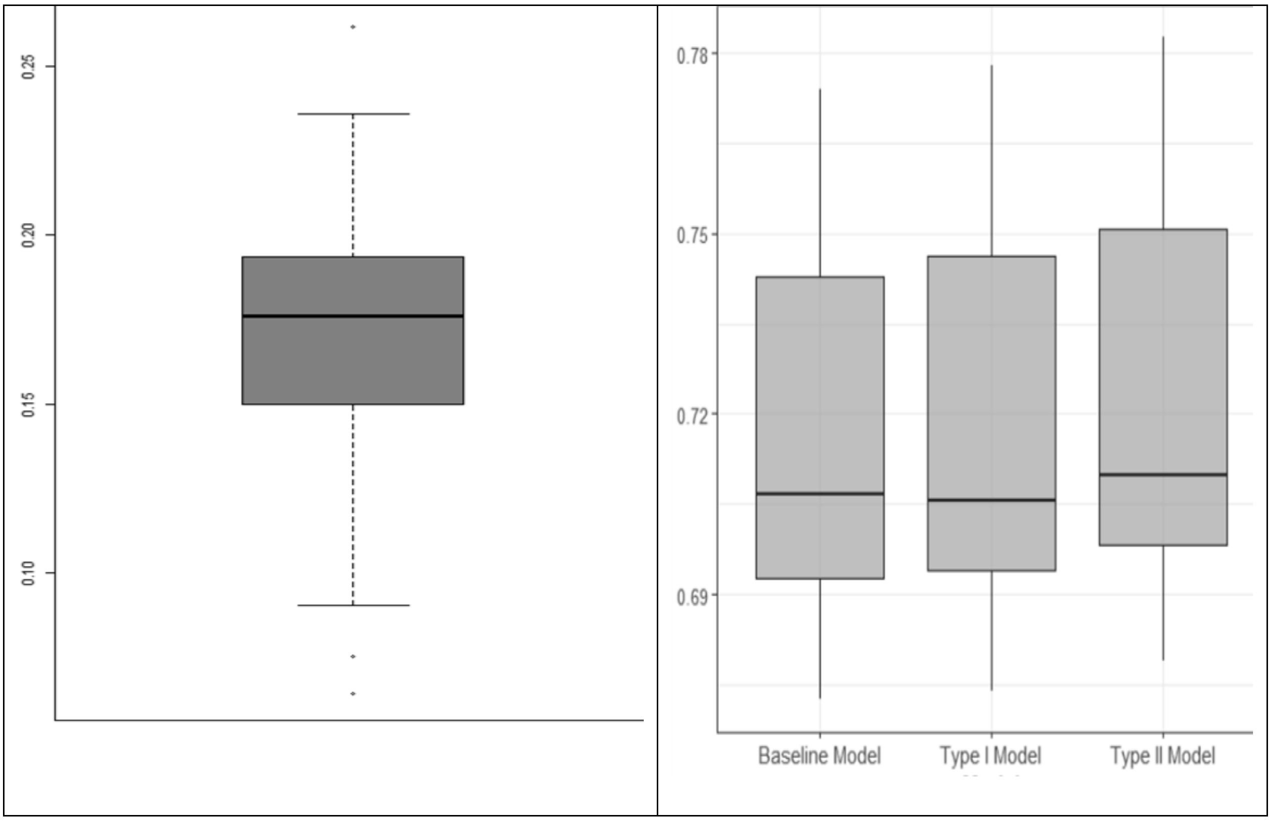

Figure 4a. (left) Box-Plot of ρ estimation in training cross validation subset.

Figure 4b. (right) Box-Plot of AUC estimated in test cross-validation subset.