PROGRAMA DECIDE

Tesis Doctoral

A Review of Spatial Probit Models: Estimation, Model Selection and Applications focused on Human Behavior

Miguel Ángel de la Llave Montiel

mdelallav5@alumno.uned.es

Director: Fernando López Hernández

Madrid, 2022

Prólogo y Agradecimientos

La econometría espacial es una ciencia en clara expansión con numerosas y heterogéneas aplicaciones. Esta Tesis Doctoral tiene como objetivo contribuir en cierta medida al desarrollo de los modelos econométricos Probit espaciales. La intención en cada uno de los capítulos ha sido el de bajar al mínimo detalle, proporcionando una mejor comprensión de los procesos fundamentales detrás de datos dicotómicos con estructura espacial.

La presente Tesis la componen cuatro capítulos independientes cuyo elemento común es la modelización Probit espacial. Realizamos una revisión completa del estado de situación de este tipo de proceso, analizamos las alternativas disponibles para su estimación y comparamos los algoritmos de estimación bajo diferentes condiciones de simulación. Se proponen dos estrategias para la selección de la verdadera forma funcional de la que proceden los datos. Revisamos las aplicaciones realizadas por investigadores en el terreno del análisis del comportamiento humano y desarrollamos dos aplicaciones novedosas en campos donde no se había aplicado estas metodologías hasta el momento.

El trabajo de investigación puede servir a futuros investigadores en la materia al igual que a empresas que basen sus decisiones de negocio en los datos y su entendimiento. Se resuelven temas no contemplados hasta el momento, por lo que se cubre un vacío importante en la materia y se identifican investigaciones adicionales necesarias para seguir avanzando. Las aplicaciones desarrolladas pueden servir de inspiración a equipos analíticos interesados en sacar buen rendimiento de los datos.

Creo que para el desarrollo de una ciencia como esta, deben confluir ciertos elementos con cierto grado de madurez como las técnicas de computación, métodos matemáticos, almacenamiento de datos. Pero quizás el elemento más importante es la comunidad científica que lo compone. Las personas definen el camino y en este camino yo me he encontrado con extraordinarios conocedores de la materia pero infinitamente mejores personas. Todo mi agradecimiento y admiración a Fernando López, catedrático de la Universidad Politécnica de Cartagena por su constante empeño en incrementar la excelencia técnica y darme la oportunidad de meterme en estas materias tan interesantes. Igualmente, destacar y agradecer a la UNED que haya hecho posible todo esto.

Esta tesis está dedicada a mi familia y amigos. Gracias a Pilar por su ayuda y ánimo constante. En estos 6 años de Tesis Doctoral hemos vivido años increibles, han nacido nuestros dos hijos Claudia y Alejandro que son lo mejor de nuestras vidas y nos recargan con ilusión. A mis padres por su incansable y constante apoyo. Al resto de mi familia y amigos que son muy grandes. A Fernando de nuevo que además de buen director, lo considero compañero y amigo y siempre es un placer hablar con él. A Ana que ayudó incondicionalmente, para que quede su recuerdo en esta Tesis. Después de 6 años creo que estas son las últimas frases de la Tesis Doctoral de la que estoy muy orgulloso.

Madrid, 8 de Marzo de 2022.

Resumen de la Tesis

Esta Tesis Doctoral la conforman cuatro estudios independientes enmarcados dentro de la temática de los modelos Probit espaciales. El objetivo que perseguimos en las siguientes páginas es el de realizar una revisión completa de los procesos dicotómicos sección cruzada con componente espacial a través de modelos Probit. A lo largo del documento mostramos los numerosos avances en Econometría espacial durante los últimos años e incorporamos novedosos estudios enfocados hacia la modelización de este tipo de procesos desde la perspectiva de la selección de modelos y estudios aplicados al análisis del comportamiento del consumidor desde la óptica espacial. Estas investigaciones muy específicas contribuyen a robustecer la ya consolidada ciencia de la Econometría espacial. A continuación, se resumen los principales puntos y contribuciones de cada uno de los capítulos esta Tesis Doctoral.

Hace más de 40 años, en Paelinck and Klaassen (1979) se acuñó el término de Econometría espacial donde se especificaron las características fundamentales de esta ciencia, cuyos métodos tratan de resolver el problema de especificación y estimación cuando el espacio o las interdependencias entre observaciones juegan un papel decisivo en la explicación de un fenómeno. Para la consolidación de este conjunto de técnicas no solo ha hecho falta el interés de una comunidad científica en búsqueda de mejores estimaciones, sino que ha tenido que venir acompañada de una creciente disponibilidad de datos georreferenciados, capacidades de computación más elevadas y diseño de algoritmos eficientes para la estimación de modelos espaciales. En el **Capítulo 1** se habla de esta evolución y los avances científicos en la materia. La realidad es que muy pronto se empezó a prestar atención a los modelos dicotómicos tan presentes en la economía (McMillen, 1992) y, de una manera más bien intermitente, han ido surgiendo métodos con los que estimar los coeficientes de un modelo Probit espacial eficientemente. La estimación del Probit espacial es un proceso complejo y requiere de técnicas avanzadas para su correcto cálculo. La interdependencia de las observaciones conlleva la no esfericidad de los residuos y dado que la probabilidad de éxito de cada observación está vinculada con la probabilidad de éxito del resto de observaciones, el modelo no puede ser resuelto como producto de las N distribuciones marginales, sino que se tendrá que maximizar el logaritmo de una distribución multivariante de dimensión N – Siendo N el tamaño muestral. No hay solución analítica para el cómputo de esta integral múltiple y hay que recurrir a otro tipo de técnicas. A lo largo del capítulo se desarrollan las metodologías propuestas para la solución del problema basándonos en los artículos originales. Por orden cronológico, el algoritmo de *Expectation-Maximization* (EM) propuesto en McMillen (1992), la *Generalización del método de los momentos* (GMM) en Pinkse and Slade (1998), los métodos basados en muestreo primero *Gibbs bayesian sampling* (Gibbs) por LeSage (2000a) y más adelante *Recursive Importance Sampling* (RIS) por Beron and Vijverberg (2004a), luego una linealización del algoritmo

GMM (LGMM) por Klier and McMillen (2008) y por último en Martinetti and Geniaux (2017) se resuelve el problema aproximando la función de verosimilitud (ML) con un método numérico.

Entre las aportaciones más relevantes de este primer capítulo se encuentra la precisión en las estimaciones de dichos algoritmos a través de un ejercicio de Monte Carlo ante un modelo autorregresivo espacial. Las conclusiones extraídas complementan el trabajo de Calabrese and Elkink (2014). En la presente tesis, se incorporan todos los algoritmos hasta la fecha, que por una cuestión temporal evidentemente no se contemplaban en Calabrese and Elkink (2014), se incorpora una visión "no-ideal" de la especificación introduciendo endogeneidad y además utilizamos los algoritmos preparados y optimizados en paquetes de R para su comparativa (Gibbs, LGMM y ML). La principal diferencia entre los algoritmos es en el sesgo de los coeficientes y en el coste computacional. EM y GMM proporcionan estimaciones pobres en cuanto a su exactitud en los parámetros. RIS, por el contrario, funciona razonablemente bien en la estimación, pero el tiempo de computación es desmesurado incluso para muestras de 400-900 observaciones. El algoritmo más rápido sin duda es LGMM, puesto que evita el problema de invertir matrices de adyacencia, lo cual rebaja mucho la carga computacional. La principal desventaja de LGMM es la infraestimación de del parámetro de dependencia espacial cuando esta es elevada y que provoca un sesgo fuerte en los coeficientes y un descenso de la precisión medida con la curva ROC. Entre Gibbs y ML no hay diferencias significativas en precisión. La metodología MCMC que utiliza Gibbs proporciona un ajuste lento hacia la solución del problema, sin embargo, después de la optimización de Wilhelm and de Matos (2015) el proceso de convergencia es más largo que ML pero dentro de unos límites razonables de tiempo.

La principal razón, identificada en Arbia (2014), por la que las investigaciones con Probit espacial hayan recibido menos atención es precisamente por la complejidad técnica del proceso de cálculo. Recientemente se ha publicado un estudio que muestra la evolución de la modelización de procesos discretos en diferentes campos (Haghani et al., 2021). La relevancia del factor espacial dentro de esta investigación es prácticamente nula. En Billé and Arbia (2019) se reivindica la necesidad de utilización de estas técnicas en el terreno de la economía de la salud. Por nuestra parte, en la última sección del capítulo hacemos una valoración del estado de estas técnicas en la modelización del comportamiento humano. Queda todo un mundo por recorrer en cuanto a la aplicación de estas técnicas a casos específicos. Sin duda el comportamiento ante la decisión sobre el uso del terreno es el tópico que más contribuciones ha recibido. Probablemente exista un efecto llamada entre investigaciones a la hora de la elección del tipo de modelización a emplear. Esto debe extrapolarse a otras áreas de estudio que necesitan más contribuciones para la explicación y valoración de fenómenos. Encontramos numerosos papers en los que se analiza el comportamiento ante la decisión de compra de un tipo de vehículo o decisiones

sobre el uso del transporte público o privado sin tener en cuenta factores espaciales o la interconexión entre individuos. Como se verá más adelante en el último capítulo de la tesis, la no selección de una correcta especificación tiene una importancia clave en la correcta interpretación de los resultados y posibles acciones detrás de los modelos econométricos. Por esta razón, con este capítulo también queremos de alguna manera reivindicar la importancia de los modelos espaciales para explicar el comportamiento humano que ya se ha tenido en cuenta en muchas investigaciones (Wang et al., 2015, Holloway et al. (2002), Arima (2016), de la Llave Montiel and López (2020) ....) y que sin duda será clave en años venideros.

La utilización del modelo Probit espacial nos estará aportando estimaciones de la estructura de correlación condicionada entre un conjunto de variables independientes y una variable dicotómica respuesta teniendo en cuenta el efecto espacial. La estimación por métodos espaciales resuelve el problema de la interconexión de observaciones, que apenas tiene que ver con la causalidad del problema (Rüttenauer, 2019) para la que habría que recurrir a técnicas experimentales o a tests específicamente diseñados para detectar la dirección de la causalidad (Herrera-Gomez et al., 2021). El **Capítulo 2** de la presente Tesis basado en (De la Llave et al., 2019b), trata de resolver el problema de fuga de clientes en una entidad de seguros a través de un modelo autorregresivo espacial. El estudio se centra en resolver principalmente la correlación entre un set de datos y la variable endógena dicotómica encontrando una forma funcional idónea y teorizamos sobre las posibles causas origen. El coeficiente autorregresivo del modelo final [0.215] nos dice que un porcentaje significativo de la probabilidad final del cliente de abandonar la compañía viene por efectos marginales indirectos. En este caso específico la causa raíz más probable es el efecto boca a boca y las experiencias familiares y vecinales. La contribución más directa de la investigación es la solución al problema de fuga con un método econométrico espacial. Numerosas son las investigaciones anuales que abordan el problema de fuga con modelos clásicos logit o con árboles de decisión (Günther et al., 2014, Lemmens and Croux (2006)..), aunque ganan protagonismo las nuevas metodologías más relacionadas con métodos no interpretables (Hung et al., 2006, Xie et al. (2009)..). Dado que es un problema ampliamente estudiado, la presente investigación busca resaltar aquellos detalles más importantes a tener en cuenta y que hasta ahora han pasado desapercibidos. Uno de los principales es la invalidez del modelo Probit clásico en el que se rechaza la hipótesis nula de que los residuos no tengan autocorrelación (Kelejian and Prucha, 2001a), y por consiguiente proporciona estimadores sesgados e inconsistentes sobre los que sería peligroso tomar decisiones. El modelo espacial final resultante del análisis muestra mejores estadísticos que el modelo clásico tanto en precisión como en idoneidad de los residuos. A partir de este modelo se aportan los efectos marginales tanto directos como indirectos de cada una de las variables. Una de las variables más notable y novedosa en este tipo de modelos es la distancia entre observaciones y lugares relevantes. Se proporciona la cuantificación de la reducción de probabilidad al establecer una sucursal cerca

de los clientes propios o por el contrario, cuando la competencia estrecha distancias con clientes ajenos la desvinculación de estos con su compañía original es evidente. Este tipo de variables no es común en este tipo de literatura probablemente por las dificultades que suponía, hasta hace unos años, conseguir información georreferenciada. Sin duda la incorporación de factores geográficos urbanos junto con especificaciones más realistas, como la propuesta en el estudio, ayudará a tomar decisiones de negocio para las empresas basadas en los datos de una manera más precisa y eficiente.

Los modelos econométricos quedan a la sombra del creciente boom de los modelos ensamblados y de redes profundas. El mundo del Big Data ha traído un creciente interés por tipos de modelización que buscan la mayor precisión posible en el resultado. Esta alta precisión en busca de la mejor combinación no lineal hace que la interpretabilidad sea un punto fuerte a desarrollar (Gilpin et al., 2018). Somos conscientes, en relación al capítulo 2, de que la precisión global del fenómeno podría mejorar utilizando técnicas del llamado "Deep-Learning"; sin embargo, nuestra propuesta es la de desmigar a fondo la información, en búsqueda de una forma funcional realista (modelo autorregresivo) y unas variables correctamente adaptadas. Para ello introducimos dentro del proceso de estimación una fase de deslinealización a través de la técnica de Multivariate Adaptive Regression Splines MARS) (Milborrow, 2011). Como se demuestra en la literatura académica, la edad es clave para entender la fuga. Pero con la técnica MARS identificamos que la pendiente de la beta no es constante para todas las edades, habiendo un punto de inflexión en 46 años. Lo mismo ocurre con la prima pagada que tiene este cambio en 549 euros.

Con el conjunto de técnicas que presentamos en el capítulo 2, cubrimos un importante vacío existente en torno a la fuga de clientes aportando nuevas ideas y una visión hasta ahora no tratadas. Aprovechando el conocimiento generado en esta investigación surge el **Capítulo 3** de la Tesis basado en de la Llave Montiel and López (2020), en la que profundizamos en un tema de indiscutible actualidad. En este capítulo indagamos en los factores que mueven al usuario de una aplicación de compra por internet a permanecer inactivo por una larga duración (>4meses) después de un primer contacto con la App. Este periodo va en línea con lo establecido investigaciones relacionadas con otro tipo de servicios (Buckinx and Van den Poel, 2005, Lai and Zeng (2014)). Sin embargo, nada hay escrito sobre la fidelización del cliente de comercio electrónico cuyo negocio está actualmente en expansión (Frasquet Deltoro et al., 2012). Una vez más el modelo Probit espacial mejora los resultados del modelo clásico y por lo tanto es una evidencia más sobre los beneficios de este tipo de modelización y su empleabilidad en la gestión de negocios. El comportamiento de la inactividad de clientes cercanos se demuestra estar codeterminado, detectándose conductas miméticas entre ellos. El número de vecinos más relevante para el modelo se calcula a través del estadístico Join-Count (Cliff and Ord, 1981), seleccionado los 5 vecinos más cercanos, que es donde a priori la autocorrelación

espacial de la variable dicotómica es más alta. Adicionalmente, testamos los residuos del modelo clásico y modelo espacial con el I-Moran generalizado con 5,10 y 15 vecinos más cercanos para robustecer los resultados. La ROC conseguida en el modelo final supera el 70%, lo cual es relevante dado que estas aplicaciones no hay mucha información personal del usuario. Como consecuencia, además de la información de inicio de sesión en la aplicación, primeros movimientos y compras realizadas, se ha añadido información geográfica muy importante. La distancia, medida en logaritmos, entre el cliente y el centro comercial más próximo es esencial para medir la desvinculación futura del cliente. A mayor distancia menor probabilidad de que el usuario se inactive por larga duración. Además, a partir del primer kilómetro parece que hay un punto de inflexión en el que la probabilidad de inactividad empieza a disminuir. En este caso, para contribuir con un nuevo aporte en la materia, se han identificado las no-linealidades a través de un modelo general aditivo (GAM) (Hastie and Tibshirani, 1986).

Los modelos con variables exógenas retardadas espacialmente cobran importancia en la literatura econométrica espacial. En Elhorst et al. (2014) y LeSage (2014a) se destaca el atractivo de este tipo de modelos por su flexibilidad, interpretabilidad y facilidad en la implementación. Dentro del modelo econométrico propuesto en el capítulo 3, tiene mucha relevancia la incorporación de este tipo de variables. Una de las variables más decisivas para predecir la inactividad del usuario es la actividad registrada los primeros días de uso de la app. El efecto marginal directo más indirecto del número de órdenes dadas durante la primera semana es superior a un 20% de reducción de probabilidad. A lo largo de la investigación, observamos que esta variable retardada espacialmente con la matriz de adyacencia de 5 vecinos, aporta un efecto marginal de un 5%. Ello quiere decir que, para el caso de un cliente que no haga operaciones con la aplicación una vez descargada, y por lo tanto su probabilidad de inactividad crezca abruptamente; en el caso de que sus vecinos si que la hayan utilizado, este incremento se rebaja sustancialmente. Por lo tanto, el modelo propuesto resultante conjugará el efecto autorregresivo y el efecto del retardo espacial de variables independientes, lo cual es un modelo econométrico por consenso muy apropiado para realizar este tipo de estudios (Rüttenauer, 2019). Podemos decir que el capítulo abre una línea de investigación para el estudio del comercio online en la que escasean los datos y las contribuciones. Sin duda este foco de análisis ayudará a empresas emergentes en comercio electrónico a buscar soluciones para la mejora de experiencia del consumidor.

Los modelos presentados en los capítulos anteriores pretenden hacer un análisis exhaustivo de cada fenómeno dicotómico. Además de ser una potente herramienta de predicción, también nos sirven para explicar el pasado y valorar el efecto de acciones/políticas empresariales. En la investigación económica es muy importante estos últimos dos puntos. Por citar algunos ejemplos, Ortega-García et al. (2017) realiza un análisis de factores que inciden en la aparición de cáncer infantil descubriendo una posible asociación entre

la aparición de la enfermedad y la exposición a la contaminación producida por ciertas industrias. LeSage et al. (2011), explora la recuperación de la actividad comercial tras la devastación del huracán Katrina en Nueva Orleans. La incorrecta identificación de la correcta especificación provoca inconsistencias y sesgos graves en los parámetros como demostramos en el **Capítulo 4** (Paper en proceso de publicación). Por lo tanto, las conclusiones o acciones derivadas de un modelo incorrectamente especificado serían más que cuestionables. La búsqueda de la correcta selección de modelos espaciales tiene escasas contribuciones por el momento. Los principales artículos se centran en modelos espaciales continuos (Florax et al., 2003; Mur and Angulo, 2009; Agiakloglou and Tsimpanos, 2021) y tan sólo encontramos uno que aborda modelos Probit comparando solamente tres tipos de modelos Beron and Vijverberg (2004b). La principal aportación del capítulo es la configuración de dos algoritmos de selección de la verdadera especificación del modelo Probit espacial. Siguiendo la discusión del modelo continuo (Florax et al., 2003; Mur and Angulo, 2009) proponemos una estrategia de lo específico a lo general (Stge) y otra estrategia de lo general a lo específico (Gets). La comparativa entre ambas técnicas la realizamos a través de una simulación de Monte Carlo para 5 tipos de especificaciones reales: Modelo Independiente (SIM), Modelo Autorregresivo Espacial (SAR), Modelo de Dependencia Espacial en el Error (SEM), Modelo con variable retardada espacialmente (SLX) y Modelo autorregresivo con variable retardada espacialmente-Durbin (SDM), 5 tamaños muestrales (100,400,900,1600,2500) y 5 parámetros de dependencia espacial (0.3,0.4,0.5,0.6,0.7). Las estrategias de selección presentan un rendimiento superior al 85% de casos correctamente seleccionados para muestras superiores a 900 observaciones. Resulta difícil decidir contundentemente qué estrategia es la mejor. Bajo condiciones ideales parece que Stge funciona ligeramente mejor que Gets. Sin embargo, cuando introducimos simulaciones con condiciones no ideales como endogeneidad en el modelo o falta de información entonces hay ciertas ocasiones que Gets es menos sensible.

En las investigaciones aplicadas, vemos como no existe un criterio homogéneo para determinar la forma funcional de un modelo Probit espacial. Desde el uso del ratio de verosimilitudes (LR) para comparar modelos (Mate-Sánchez-Val, 2021) hasta seleccionar el modelo que mejor precisión ofrezca (Läpple et al., 2017). Nuestro capítulo demuestra que al combinar diferentes tests [I-Moran Generalizado, t-test, LR test, LR Confac] podemos conseguir resultados bastante atractivos que determinen la procedencia de los datos. Las estrategias mostradas siguen el principio de parsimonia y tienen un sentido econométrico estricto en cada paso. Reservamos una serie de tests que complicarían las estrategias más de lo necesario [Join Count, AIC, BIC]. La solución que damos de que las estrategias están cerca de un óptimo con toda la batería de tests disponibles es aplicando un algoritmo Gradient Boosting (GBM) de clasificación multinivel. La conclusión es que no se aprecian diferencias notables entre las estrategias aportadas y el resultado del GBM para muestras superiores a 400 observaciones.

Cada uno de los capítulos presentados pretende continuar el estado de conocimiento dentro del campo de la Econometría espacial. Con toda probabilidad, esta ciencia seguirá avanzando en cuestiones metodológicas, aplicaciones prácticas, desarrollo de nuevos tests. Hemos desarrollado esta Tesis para responder a algunas preguntas que quedaban pendientes y servirá como punto de apoyo para futuras investigaciones.

Summary of the Thesis

This Thesis contains four independent studies framed within the theme of spatial Probit models. The objective that we pursue in the following pages is to carry out a complete review of dichotomous cross-section processes with a spatial component through Probit models. Throughout the document we show the numerous advances in Spatial Econometrics in recent years and we incorporate new studies focused on the modeling of this type of process from the perspective of model selection and studies applied to the analysis of consumer behavior from the spatial perspective. These very specific investigations contribute to strengthen the already consolidated science of Spatial Econometrics. The main points and contributions of each of the chapters of this Thesis are summarized below.

More than 40 years ago, Paelinck and Klaassen (1979) coined the term Spatial Econometrics where the fundamental characteristics of this science were specified, whose methods try to solve the problem of specification and estimation when space or the interdependencies between observations play a decisive role in the explanation of a phenomenon. The consolidation of this set of techniques has not only required the interest of a scientific community in search of better estimates, but it has also had to be accompanied by a growing availability of georeferenced data, higher computing capacities and the design of efficient algorithms to estimate spatial models. **Chapter 1** explains this evolution and the scientific advances in the topic. The reality is that very soon, attention began to be paid to the dichotomous models so present in the economy (McMillen, 1992) and, rather intermittently, methods have emerged with which to estimate the coefficients of a spatial Probit model efficiently. The estimation of the spatial Probit is a complex process and requires advanced techniques for its correct calculation. The interdependence of the observations entails the non-sphericity of the residuals and since the probability of success of each observation is linked to the probability of success of the rest of the observations, the model cannot be solved as a product of the N marginal distributions, but rather you will have to maximize the logarithm of a multivariate distribution of dimension N – N being the sample size. There is no analytical solution for the computation of this multiple integral and it is necessary to resort to other types of techniques. Throughout the chapter, the proposed methodologies for solving the problem are developed based on the original articles. In chronological order, the *Expectation-Maximization* (EM) algorithm proposed in McMillen (1992), the *Generalization of Moments Method* (GMM) in Pinkse and Slade (1998), *Gibbs bayesian sampling* (Gibbs) first sampling-based methods by LeSage (2000a) and later *Recursive Importance Sampling* (RIS) by Beron and Vijverberg (2004a), then a linearization of the GMM algorithm (LGMM) by Klier and McMillen (2008) and finally in Martinetti and Geniaux (2017) the problem is solved by approximating the likelihood function (ML) with a numerical method.

Among the most relevant contributions of this first chapter is the precision in the

estimations of the algorithms presented through a Monte Carlo exercise for a spatial autoregressive model. The conclusions drawn complement the work of Calabrese and Elkink (2014). In this thesis, all the algorithms to date are incorporated in the study, which for a temporary reason were evidently not contemplated in Calabrese and Elkink (2014). A "non-ideal" vision of the specification is incorporated by introducing endogeneity and we also use the algorithms prepared and optimized in R packages for comparison (Gibbs, LGMM and ML). The main difference between the algorithms is in the bias of the coefficients and in the computational cost. EM and GMM provide poor estimates in terms of their parameter accuracy. RIS, on the other hand, performs reasonably well at estimation, but computation time is very high even for samples of 400-900 observations. The fastest algorithm is undoubtedly LGMM, since it avoids the problem of inverting adjacency matrices, which greatly reduces the computational time. The main disadvantage of LGMM is the underestimation of the spatial dependence parameter when it is high, which causes a strong bias in the coefficients and a decrease in the precision measured with the ROC curve. Between Gibbs and ML there are no significant differences in precision. The MCMC methodology used by Gibbs provides a slow adjustment towards the solution of the problem, however after Wilhelm and de Matos (2015) optimization the convergence process is longer than ML but within reasonable time limits.

The main reason, identified in Arbia (2014), why spatial Probit research has received less attention is precisely because of the technical complexity of the calculation process. A recent study published shows the evolution of discrete process modeling in different fields (Haghani et al., 2021). The relevance of the spatial factor in this research is barely significant. In Billé and Arbia (2019) the need to use these techniques in the field of health economics is claimed. For our part, in the last section of the chapter we make an assessment of the state of these techniques in the modeling of human behavior. There is still a whole world to go in terms of the application of these techniques to specific cases. Undoubtedly, the behavior before the decision on the use of the land is the topic that has received the most contributions. There is probably a pull effect between investigations when choosing the type of modeling to use. This must be extrapolated to other areas of study that need more contributions for the explanation and assessment of phenomena. We find numerous papers in which behavior is analyzed when deciding to purchase a type of vehicle or decisions about the use of public or private transport without taking into account spatial factors or the interconnection between individuals. As will be seen later on, in the last chapter of the thesis, the non-selection of a correct specification has a key importance in the correct interpretation of the results and possible actions behind the econometric models. For this reason, with this chapter we also want in some way to claim the importance of spatial models to explain human behavior that has already been taken into account in many investigations (Wang et al., 2015, Holloway et al. (2002), Arima (2016), de la Llave Montiel and López (2020) ....) and that will undoubtedly be key in years to come.

The use of the spatial Probit model will be providing us with estimates of the conditional correlation structure between a set of independent variables and a dichotomous response variable, taking into account the spatial effect. Estimation by spatial methods solves the problem of the interconnection of observations, which has little to do with the causality of the problem (Rüttenauer, 2019) for which it would be necessary to resort to experimental techniques or tests specifically designed to detect the direction of causality (Herrera-Gomez et al., 2021). **Chapter 2** of this Thesis based on (De la Llave et al., 2019b), tries to solve the problem of customer churn in an insurance entity through a spatial autoregressive model. The study focuses mainly on solving the correlation between a data set and the dichotomous endogenous variable by finding a suitable functional form and theorizing about the possible root causes. The autoregressive coefficient of the final model [0.215] tells us that a significant percentage of the client's final probability of leaving the company comes from indirect marginal effects. In this specific case, the most likely root cause is word of mouth and family and neighborhood experiences. The most direct contribution of the research is the solution to the leakage problem with a spatial econometric method. There are numerous annual investigations that address the leakage problem with classic logit models or with decision trees (Günther et al., 2014, Lemmens and Croux (2006)..), although new methodologies more related to non-interpretable methods (Hung et al., 2006, Xie et al. (2009)..) gain prominence. Since it is a widely studied problem, this research seeks to highlight those details that are most important to take into account and that until now have gone unnoticed. One of the main ones is the invalidity of the classical Probit model in which the null hypothesis that the residuals have no autocorrelation (Kelejian and Prucha, 2001a) is rejected, and therefore provides biased and inconsistent estimators on which it would be dangerous to make decisions. The final spatial model resulting from the analysis shows better statistics than the classical model both in terms of accuracy and suitability of the residuals. From this model, the direct and indirect marginal effects of each of the variables are provided. One of the most notable and novel variables in this type of model is the distance between observations and relevant places. The quantification of the reduction in probability is provided when establishing a branch near one's own clients or, on the contrary, when the competition closes distances with foreign clients, the disconnection of these with their original company is evident. This type of variable is not common in this type of literature, probably due to the difficulties involved, until a few years ago, in obtaining georeferenced information. Without a doubt, the incorporation of urban geographic factors together with more realistic specifications, such as the one proposed in the study, will help make business decisions for companies based on data in a more precise and efficient way.

Econometric models are overshadowed by the growing boom in assembled and deep network models. The world of Big Data has brought a growing interest in types of modeling that seek the highest possible precision in the result. This high precision in search of the best non-linear combination makes interpretability a strong point to develop (Gilpin

et al., 2018). We are aware, in relation to chapter 2, that the global precision of the phenomenon could be improved using techniques called "Deep-Learning"; however, our proposal is to thoroughly dissect the information, in search of a realistic functional form (autoregressive model) and correctly adapted variables. To do this, we introduce a de-linearization phase within the estimation process through the Multivariate Adaptive Regression Splines (MARS) (Milborrow, 2011) technique. As demonstrated in the academic literature, age is key to understanding fugue. But with the MARS technique we identify that the beta slope is not constant for all ages, with a turning point at 46 years. The same happens with the paid premium that has this change at 549 euros.

With the set of techniques that we present in Chapter 2, we fill an important gap around customer churn by providing new ideas and insights hitherto untouched. Taking advantage of the knowledge generated in this research, **Chapter 3** of the Thesis based on de la Llave Montiel and López (2020) arises, in which we delve into an indisputably current topic. In this chapter we investigate the factors that move the user of an online shopping application to remain inactive for a long period of time (>4 months) after a first contact with the App. This period is in line with what has been established in research related to another type of services (Buckinx and Van den Poel, 2005, Lai and Zeng (2014)). However, nothing is written about the loyalty of the e-commerce customer whose business is currently expanding (Frasquet Deltoro et al., 2012). Once again, the spatial Probit model improves the results of the classical model and is therefore further evidence of the benefits of this type of modeling and its employability in business management. The behavior of nearby clients is shown to be co-determined, detecting mimetic behaviors between them. The most relevant number of neighbors for the model is calculated through the Join-Count statistic (Cliff and Ord, 1981), selecting the 5 closest neighbors, which is where the spatial autocorrelation of the dichotomous variable is higher. Additionally, we test the residuals of the classical model and spatial model with the generalized I-Moran with 5, 10 and 15 nearest neighbors to strengthen the results. The ROC achieved in the final model exceeds 70%, which is relevant given that these applications do not contain much personal user information. As a consequence, in addition to the login information in the application, first movements and purchases made, very important geographical information has been added. The distance, measured in logarithms, between the customer and the nearest shopping center is essential to measure future customer disengagement. The greater the distance, the less likely the user will be inactive for a long time. In addition, from the first kilometer it seems that there is a turning point in which the probability of inactivity begins to decrease. In this case, to contribute a new contribution to the matter, non-linearities have been identified through a general additive model (GAM) (Hastie and Tibshirani, 1986).

Models with spatially lagged exogenous variables gain importance in the spatial econometric literature. Elhorst et al. (2014) and LeSage (2014a) highlight the attractiveness

of this type of model due to its flexibility, interpretability and ease of implementation. Within the econometric model proposed in chapter 3, the incorporation of this type of variables is very important. One of the most decisive variables to predict user inactivity is the activity registered the first days of use of the app. The direct plus indirect marginal effect of the number of orders given during the first week is greater than a 20% probability reduction. Throughout the investigation, we observed that this spatially lagged variable with the 5-neighbor adjacency matrix provides a marginal effect of 5%. This means that, in the case of a client that does not perform operations with the application once downloaded, and therefore its probability of inactivity increases abruptly; in the event that their neighbors have used it, this increase is substantially reduced. Therefore, the resulting proposed model will combine the autoregressive effect and the effect of the spatial lag of independent variables, which is a very appropriate consensus econometric model to carry out this type of study (Rüttenauer, 2019). We can say that the chapter opens a line of research for the study of online commerce in which data and contributions are scarce. Without a doubt, this focus of analysis will help emerging companies in electronic commerce to look for solutions to improve the consumer experience.

The models presented in the previous chapters aim to make an exhaustive analysis of each dichotomous phenomenon. In addition to being a powerful prediction tool, they also serve to explain the past and assess the effect of business actions/policies. In economic research, these last two points are very important. To cite a few examples, Ortega-García et al. (2017) performs an analysis of factors that affect the appearance of childhood cancer, discovering a possible association between the appearance of the disease and exposure to pollution produced by certain industries. LeSage et al. (2011), explores the recovery of business after the devastation of Hurricane Katrina in New Orleans. The incorrect identification of the correct specification causes inconsistencies and serious biases in the parameters as we demonstrate in **Chapter 4** (Paper in publication process). Therefore, the conclusions or actions derived from an incorrectly specified model would be more than questionable. The search for the correct selection of spatial models has few contributions at the moment. The main articles focus on continuous spatial models (Florax et al., 2003; Mur and Angulo, 2009; Agiakloglou and Tsimpanos, 2021) and we only found one that addresses Probit models by comparing only three types of models Beron and Vijverberg (2004b). The main contribution of the chapter is the configuration of two selection algorithms of the true specification of the spatial Probit model. Following the discussion of the continuous model (Florax et al., 2003; Mur and Angulo, 2009) we propose a strategy from the specific to the general (Stge) and another strategy from the general to the specific (Gets). The comparison between both techniques is carried out through a Monte Carlo simulation for 5 types of real specifications: Independent Model (SIM), Spatial Autoregressive Model (SAR), Spatial Error Dependence Model (SEM), Model with lagged variable spatially (SLX) and Autoregressive Model with spatially lagged variable-Durbin (SDM), 5 sample sizes (100,400,900,1600,2500) and 5 parameters

of spatial dependence (0.3,0.4,0.5,0.6,0.7). The selection strategies present a performance greater than 85% of correctly selected cases for samples greater than 900 observations. It is difficult to decide decisively which strategy is the best. Under ideal conditions Stge seems to perform slightly better than Gets. However, when we introduce simulations with non-ideal conditions such as endogeneity in the model or lack of information, then there are certain occasions that Gets is less sensitive.

In applied research, we see how there is no homogeneous criterion to determine the functional form of a spatial Probit model. From the use of the likelihood ratio (LR) to compare models (Mate-Sánchez-Val, 2021) to selecting the model that offers the best accuracy (Läpple et al., 2017). Our chapter shows that by combining different tests [Generalized I-Moran, t-Test, LR Test, LR Confac] we can get quite attractive results that determine the origin of the data. The strategies shown follow the parsimony principle and make strict econometric sense at each step. We reserved a series of tests that would complicate the strategies more than necessary [Join Count, AIC, BIC]. The solution we give that the strategies are close to an optimum with the entire battery of available tests is by applying a multilevel classification Gradient Boosting (GBM) algorithm. The conclusion is that there are no notable differences between the strategies provided and the result of the GBM for samples greater than 400 observations.

Each of the chapters presented aims to continue the state of knowledge within the field of Spatial Econometrics. In all likelihood, this science will continue to advance in methodological issues, practical applications, development of new tests. We have developed this Thesis to answer some questions that remained pending and will serve as a point of support for future research.

List of Figures

List of Tables

Contents

Chapter 1

Spatial probit models

1.1 Introduction

When we look around, we see an interconnected society. We live in family circles, in neighboring communities, in areas consistent with our lifestyle. We work and study surrounded by people with concerns similar to ours. Although technology advances fast and provides new forms of communication, cultural traditions still endure. Social, family and professional networks are present in decision-making. There are many areas interested in modeling the social and economic phenomena. Linear models over the years have become references to draw conclusions from the apparent randomness that surrounds us. These classical econometric models assume, among other key hypotheses, independence of the observations. If such independence is breached, then econometric models would provide inconsistent and inefficient estimations (McMillen, 1992). Given the existence of a certain spatial relationship between the observations that conditions the dependent variable of a model, then the need to go towards models that include spatial aspects would be evidenced. Moreover, the new technological era allows us to store and manage new information with our computers. Geospatial information has had a great increase of interest in the last decades. Georeferencing has been gaining popularity among researchers. These are some compelling reasons why during the last decades of the 20th century, spatial econometrics suffered an explosion in terms of published papers and specialized books (Anselin, 1988; Anselin and Florax, 2012; Cressie, 2015). It was the beginning of a new field in quantitative economy that moved rapidly to other social sciences. Now, after almost 40 years, spatial econometrics has reached a good level of maturity (Anselin, 2010). The scientific basis is well established, the properties of the estimators of the standard models have been studied and tests for specification and spatial autocorrelation have been developed over the years.

One topic within spatial econometrics that has received less attention are models where the dependent variable is limited. However, models with binary response are frequently

used in research in economy. When the making of a decision or the occurrence of an event follows a Bernoulli process and there is also a certain spatial dependence between observations, then logistic spatial regression is appropriate to model this behavior. The difficulty of obtaining consistent and efficient estimators meant that spatial probit and logit did not receive as much attention. At a slow but steady pace during the 90s, different mathematical techniques were developed to find solution of the problem, especially focused on the spatial probit. Additionally, the computational costs to calculate the estimators is very high, so the optimization of the algorithms has also been a key element in the development of the research. Given the advances in the theoretical part and in computer technology, it has been possible to study a great deal of topics. Here is a totally arbitrary selection of some of the works carried out. Over the last few years there have been studies focused on analyzing the generation of value through the use of the land (McMillen and McDonald, 2002), the reopening of commercial businesses after Hurricane Katrina (LeSage et al., 2011), the contagion effect in the banking crises of the late 90s in Asia (Amaral et al., 2014), the policy enrollment of different schools in Ohio (Brasington et al., 2016), mimetic spatial churn behaviors in online retail services (de la Llave Montiel and López, 2020), the probability of business failure in SMEs (Rodríguez et al., 2016). However, there are still areas where the application of this type of algorithm is not very widespread and can be very useful. This is the case of health economic issues. Billé and Arbia (2019) identify this area with a low diffusion of this type of models. Given the importance that it might have, the authors stress the need to popularize these algorithms within this type of field.

We believe that spatial econometrics will bring us good and valuable insights to understand the world. The efficiency of processes and policies must be guided by rigorous quantitative studies. This is only possible if the statistical models are sufficiently rigorous and appropriate. Throughout the following chapters of this Thesis, we carry out an in-depth analysis of the spatial probit model. Our intention is to continue contributing to the foundations of this new spatial field that emerged 40 years ago.

1.2 Spatial Probit Models. The Origins

To establish an accurate historic framework, let us now reflect briefly on the origins of the logistic regression. As Cramer (2002) brilliantly presents it, the invention of the logistic function, as well as the logit and probit models has been a long and winding road, as the Beatles would put it. It was at the end of the 18th century that the logistic function was invented, to explain the growth of population. Although Malthus (1798) was the first one to establish the exponential growth of population, it was a few years later when the conclusion that exponential growth must find resistance at some point (Verhulst, 1838), therefore introducing a term as a representation thereof. Velhurst had time to name

the function *courbe logistique*, but died at a young age, not being able to gain enough credit for his contributions to statistics. Still, sometimes randomness plays on the side of science. A few years later, such function was independently rediscovered by two other researchers (Pearl and Reed, 1920), while predicting the population growth of the United States. Pearl and Reed failed to recognize Velhurst's work and did not use the term logistic. Who did use it and address all previous work on the matter was Yule (1925). Such function is still used today to model quite disparate areas, from its original purpose – population growth, to marketing. Let us now turn our attention to the roots of the probit model. Although it is commonly attributed to Gaddum (1933) and Bliss in (Bliss, 1934), there is some evidence of a prototype of this method in Germany in the middle 19th century. According to Cramer (2002), Fechner observed differences in the human response to the same stimulus and was the first one to represent said differences in normal deviates. However, it was Gaddum and Bliss who portray the normal distribution as commonplace and establish the use of the logarithmic transformation. It was Bliss who coined the term probit, as short for *probability unit*, and who also set the *maximum likelihood* estimation for the model. The use of the probit model soon expanded: market research, economics, etc. Such model was considered superior to others, as we will see in a moment, since it was linked to the normal distribution curve. In 1944, Joseph Berckson, a physicist and statistic prone to creating controversy, who had been co-author to one of Reed's papers on autocatalytic functions, proposed the use of the logistic function, naming this model *logit*. His controversialist nature, together with the fact that he openly rejected the maximum likelihood method in favor of the chi – squared estimation, turned other scholars against him and his proposal. Despite having the academics against it, the logit method quickly spread in the actual work environment, and by the end of the 20th century both models were equally used in all areas of the social sciences. Many authors advocated for the logit model, from Cox (1969) to Theil (1969). Furthermore, in the year 2000 McFadden was awarded with a Nobel Prize for his works on the logit model. The term *logistic regression* was created to define analysis with binary dependent variable. Once computers and maximum likelihood calculator packages were widely available, logit and probit models became and have since been business as usual, both in the academic and work environment. Nowadays, probit models are probably one of the most common ways along with logit models to solve problems where the response variable is binary (Greene, 2012).

In 1979 Paelinck and Klaassen (1979) published *Spatial Econometrics*, highlighting the relevance of space locations and spatial dependence in econometrics models. That year became a turning point in this area, as research, methods, and use cases started to grow rapidly to this day that it has been reached a significant degree of maturity (Anselin, 2010). At the beginning of the era all research was much focused on standard linear regression models and tests to detect spatial dependency on model errors. However, the applications of probit models to the spatial context were not long in coming. The first

attempt to deal with the problem was in Avery et al. (1983) with a proposal of a Maximum Likelihood method to obtain estimators but assuming errors to be orthogonal to explicative variables. McMillen (1992) raised this problem given that in economic aspects heteroskedasticity is implicit in a multitude of phenomena, "if data vary spatially, it is reasonable to think that variances may vary also". McMillen proposes first solutions when heteroskedasticity and autocorrelation are observed. Since then, quite a few studies using probit method have been emerging and different ways of obtaining consistent estimators with increasingly lower computational costs have been developed. These methods will be reviewed in the 4th part of this document. On the other hand, logit spatial models seem intractable according to Anselin (2002) and therefore have received much less attention in recent years.

Throughout this chapter we are going to review from the classic probit model to the most common spatial probit models.

1.2.1 Standard probit models

Probit models arise to solve regression models where the dependent variable (Y^*) is continuous and unobservable called *latent variable* and to which a binary and observable response (Y) is associated such that:

$$\begin{aligned} Y^* &= X\beta + e \\ e &\sim N(0, \sigma^2 I) \end{aligned} \tag{1.1}$$

and

$$Y := \begin{cases} 1 & \text{if } Y^* > 0, \\ 0 & \text{if } otherwise \end{cases}] \tag{1.2}$$

Then, let (Y) be a binary Nx1 vector that reflects information that can be summarized in 0 when some event is failure and 1 for success. (Y) will be assumed to be explained linearly by a set of explicative variables (X). And the function that links the explicative variables (X) to the dependent variable (Y) is the cumulated density function of a standardized normal.

$$P(y_i = 1/x_i) = P(Y^* > 0/x_i) = P(x_i\beta + e > 0/x_i) = P(e < x_i\beta/x_i) = \\ = \Phi(x_i\beta) = \int_{-\infty}^{x_i\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \tag{1.3}$$

where Φ refers to the cumulated density function of the normal distribution and β refers to the parameters of the regression to be estimated.

contrary,

$$P(y_i = 0/x_i) = P(Y^* < 0/x_i) = P(x_i\beta + e > 0/x_i) = P(e > x_i\beta/x_i) = \\ = 1 - \Phi(x_i\beta) = 1 - \int_{-\infty}^{x_i\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \tag{1.4}$$

Maximum likelihood was the original procedure to solve it by Bliss (1935) following R. Fisher's classical book "Statistical Methods for Research Workers" and nowadays is probably the most common method to solve probit model. Since the endogenous variable is binary and there is a sample of n independent instances, the likelihood function is set as

$$L(\beta) = \prod_{y_i=1} \Phi(x_i\beta) \prod_{y_i=0} 1 - \Phi(x_i\beta) \tag{1.5}$$

wich could be expressed as:

$$L(\beta/x_i) = \prod_{i=1}^{n} (\Phi(x_i\beta))^{y_i} (1 - \Phi(x_i\beta))^{1-y_i} \tag{1.6}$$

And if natural logarithm is applied then:

$$l(\beta/x_i) = \sum_{i=1}^{n} (y_i ln(\Phi(x_i\beta)) + (1 - y_i)(1 - \Phi(x_i\beta))) \tag{1.7}$$

In order to solve the equation, the gradient of the log-likelihood function needs to be calculated and it is called the score function and it is equaled to 0.

$$l'(\beta/x_i) = \frac{\partial l(\beta/x_i)}{\partial \beta} = \sum_{i=1}^{n} \frac{y_i - \Phi(x_i\beta)}{\Phi(x_i\beta)(1-\Phi(x_i\beta))} x_i \Phi'(x_i\beta) = 0 \tag{1.8}$$

where

$$\Phi'(x_i\beta)$$

is the density function of the normal distribution. And the second partial derivative is:

$$l''(\beta/x_i) = \frac{\partial^2 l(\beta/x_i)}{\partial^2 \beta} = -\sum_{y_i=1}^{n} \lambda_i(\lambda_i + x_i\beta)x_i x_i^t \tag{1.9}$$

where λ_i is

$$\lambda_i = \frac{q_i(\Phi'(q_i x_i \beta))}{\Phi(q_i x_i \beta)} \tag{1.10}$$

where

$$q_i = (2y_i - 1)$$

Since normal cumulated density equation has no closed form, a numeric approach is needed to obtain the maximization. Newton-Raphson and Iteratively reweighted least squares are the most common methods to obtain this solution. Both are as follows:

$$Newton - Raphson: \ \beta_{t+1} = \beta_t - l''(\beta/x_i)^{-1} l'(\beta/x_i) \tag{1.11a}$$

$$Iteratively\ Reweighted\ Least\ Squares: \ \beta_{t+1} = (X^t W_t X)^{-1} X^t W_t [X\beta_t + W_t^{-1} X] \tag{1.11b}$$

where W is a diagonal matrix whose elements in the diagonal are

$$W_i = \lambda_{it}(x_i \beta_t \lambda_{it})$$

.

Regarding the interpretation of the parameters in single probit is quite easy with just the peculiarity that the marginal effect of a unitary increase of the exogenous variable on the probability estimated is not constant because it changes depending on the value of X.

$$\frac{\partial E(Y/X)}{\partial X} = \Phi'(X\beta)\beta \tag{1.12}$$

Although a common practice is to calculate the marginal effects based on the mean of all explicative variables.

1.2.2 Spatial probit models

Spatial probit models arise to extend the classical probit model by introducing some kind of spatial dependence. The outcome of the model not only depends on the characteristics of each observation, such as in the classical model, but it also relates observations closed to each other. The spatial structure can arise both by spatial dependence of the dependent variable itself or spatial dependence immersed in the errors of the model (LeSage and Pace, 2009a).

As in the classic spatial probit, let (Y*) be a binary Nx1 continuous stochastic vector of the latent variable. The broader specification would cover all types of spatial dependency. This model is the general nonlinear nesting model (GNNM)

$$\begin{aligned} Y^* &= X\beta + \rho W_1 Y^* + Z\theta + u \\ u &= \lambda W_2 u + e \\ e &\sim N(0, \sigma^2 I) \end{aligned} \tag{1.13}$$

where X is a matrix containing the initial exogenous variables and Z the same variables spatially lagged $Z = W_1 X$. Then, W_1 and W_2 are the adjacency matrices by which the observations are linked to each other. Therefore,

$$w_{ij}^* := \begin{cases} 1 & \text{If j represents one of the neighbours to i,} \\ 0 & \text{If i=j or j is not one of the neighbours to i,} \end{cases}] \tag{1.14}$$

The final W matrices are commonly row-standarized, so that

$$w_{ij} = \frac{w_{ij}^*}{\sum_{j=1}^{n} w_{ij}^*} \tag{1.15}$$

which means that W matrices become frequently asymmetric. However, there are many other specifications for weighted matrices. See Chasco (2003).

The set of coefficients to be estimated are β for each of the independent variables and their spatially lagged and also (ρ , λ, θ) which refer to the autoregressive parameters. And finally, Y is the observable variable defined by

$$Y := \begin{cases} 1 & \text{if } Y^* > 0, \\ 0 & \text{if } otherwise \end{cases}] \tag{1.16}$$

The disturbance of the model (e) is a multivariate normal variable with mean equals to 0 and finite variance, although in many theoretical research is established as 1.

Before starting with the most popular specific spatial models, it is important to highlight the assumptions made in Kelejian and Prucha (2010) in the spatial contiguity matrices and in the autoregressive parameters. The first assumption is related to the main diagonal of W{1,2} which is supposed to be 0 in all the instances. It means that no observation is a neighbor of itself. Second assumption states that the autoregressive parameters should be in the interval $(-1/\tau, 1/\tau)$. Where τ is the spectral radius of the adjacency matrices (W). The third assumption reflects that the aggregation of rows and columns of W_1, W_2, $(I - \rho W_1)^{-1}$ and $(I - \lambda W_2)^{-1}$ are bounded uniformly in absolute value. Finally, the fourth hypothesis establishes X with full rank and (X'X) nonsingular.

By setting any of the spatial parameters to zero (ρ , λ, θ), we arrive at the specific models that have been studied in the academic literature. Next, we will study the specification of these models.

1.2.2.1 Probit-SAR Model

If θ and λ are set to 0, then we obtain the so-called spatial autoregressive probit model (SAR). In this model initially presented by Anselin (1988), the endogenous variable is influenced by independent variables (X) and also depends on the lagged variable (WY*), that is, on the value of the dependent variable in neighboring locations.

$$\begin{aligned} Y^* &= X\beta + \rho W_1 Y^* + u \\ u &\sim N(0, \sigma^2 I) \end{aligned} \tag{1.17}$$

So that, the dependent variable (Y*) located in (i), can be explained by exogenous variables located in (i) and located in other locations through the spatial multiplier

$$(I - \rho W_1)^{-1}$$

, as it is reflected in next equation

$$\begin{aligned} Y^* = (I - \rho W_1)^{-1} X\beta + (I - \rho W_1)^{-1} u \\ u \sim N(0, \sigma^2 I) \end{aligned} \tag{1.18}$$

In this model, a spatial process is obtained by which in the case that Y *> 0 and therefore Y = 1 then the probability that neighboring observations have value 1 would increase.

1.2.2.2 Probit-SEM Model

In this case, if θ and ρ are set to 0, then we obtain the so called spatial error probit model (SEM). This model has been widely used in research studies in the case of Gaussian regression, unlike for probit regression. In this type of model, they assume that there are variables with a certain spatial autocorrelation and correlated with the dependent variable, and yet they are not contemplated in the model specification. Therefore, this omission of variables in the specification goes to model residuals.

$$\begin{aligned} Y^* &= X\beta + u \\ u &= \lambda W_2 u + e \\ e &\sim N(0, \sigma^2 I) \end{aligned} \tag{1.19}$$

So that, the dependent variable (Y*) located in (i), can be explained by exogenous variables located in (i) and the residuals of adjacent observations through the multiplier

$$(I - \lambda W_2)^{-1}$$

.

$$\begin{aligned} Y^* = X\beta + (I - \lambda W_2)^{-1} e \\ e \sim N(0, \sigma^2 I) \end{aligned} \tag{1.20}$$

1.2.3 A taxonomy of spatial probit models

The previous models presented (SAR and SEM) are the ones that first emerged in the spatial econometrics literature and the most widely used in research. Both models present global spillovers, as the autoregressive parameters affect the model globally. However, by

starting from the general model we can arrive at many more different models, which are those revealed by Florax and Folmer (1992).

The first specification worth addressing is the so-called *Spatial Lag of X probit model* (SLX). It is based on setting λ and ρ to zero. According to Elhorst and Halleck Vega (2017), these types of models have received very little attention over the years. However, they do deserve it since if a good specification is established with a good W adjacency matrix, it can produce very significant flexible spillovers. These spillovers in this case are considered local, given that since the diffusion effect of the spatial multiplier is less than in global cases (Chasco, 2003).

$$\begin{aligned} Y^* &= X\beta + Z\theta + u \\ u &\sim N(0, \sigma^2 I) \end{aligned} \tag{1.21}$$

Where Z is the spatial lag of variable X. We would find this type of model in a probit environment when the probability that Y = 1 in the locality (i) not only depends on the characteristics of said locality but also depends on the characteristics of the neighboring localities.

The next model to present is the one that only sets $\lambda = 0$. This would be the so-called *spatial Durbin probit model* (SDM). This model would be a combination of the previously presented SAR and SLX models, which means that combines global and local spillovers. This model introduced in LeSage and Pace (2009a) takes the following form.

$$\begin{aligned} Y^* &= X\beta + \rho W_1 Y^* + Z\theta + u \\ u &\sim N(0, \sigma^2 I) \end{aligned} \tag{1.22}$$

which can be expressed as,

$$\begin{aligned} Y^* &= (I - \rho W_1)^{-1} X\beta + (I - \rho W_1)^{-1} Z\theta + (I - \rho W_1)^{-1} u \\ u &\sim N(0, \sigma^2 I) \end{aligned} \tag{1.23}$$

In SDM the probability that Y*>0 in location (i) depends on features at location (i), also depends on features located in other locations through the spatial multiplier

$$(I - \rho W_1)^{-1}$$

, and spatial lagged features on the points surrounding (i) with the same spatial multiplier.

In case SAR and SEM models are combined, then we obtain the *spatial autoregressive model with autoregressive disturbances* (SARAR). In this model, the parameter θ is equal to zero.

$$\begin{aligned} Y^* &= X\beta + \rho W_1 Y^* + u \\ u &= \lambda W_2 u + e \\ e &\sim N(0, \sigma^2 I) \end{aligned} \tag{1.24}$$

This is a very interesting model because of the nuances it contains. A considerable aspect to note is that in this type of model the adjacency matrices W_1 and W_2 should not be exactly the same to avoid parameter estimation problems (Anselin, 1988). The model can be written also as

$$\begin{aligned} Y^* &= (I - \rho W_1)^{-1}(X\beta + (I - \lambda W_2)^{-1} e) \\ e &\sim N(0, \sigma^2 I) \end{aligned} \tag{1.25}$$

The last remarkable combination of space models is bringing together the SLX and SEM models. This would lead us to set the rho parameter to zero and give rise to the so-called model*Spatial Lag of X probit model with autoregressive disturbances.*

$$\begin{aligned} Y^* &= X\beta + Z\theta + u \\ u &= \lambda W_2 u + e \\ e &\sim N(0, \sigma^2 I) \end{aligned} \tag{1.26}$$

which can be also written as

$$\begin{aligned} Y^* &= X\beta + Z\theta + (I - \lambda W_2)^{-1} e \\ e &\sim N(0, \sigma^2 I) \end{aligned} \tag{1.27}$$

To finalize the taxonomy of the main spatial models, it can be observed how the introduction of different second-order spatial adjacency matrices would change the specification of the model. The spatial effect of spillovers would be enhanced. Although, this would undoubtedly add an additional degree of difficulty to the subject.

1.2.4 Other Alternatives:

There are many more specifications of probit models than those presented so far. Depending on the spatial lags considered, the global or local effect of the spillovers and the shape of the adjacency matrix, a multitude of different models will be achieved. The purpose of this section is to list some studies that go outside from traditional lines within spatial econometrics.

There are some investigations which deal with the problem of heteroscedasticity in spatial probit models (McMillen, 1992, and LeSage (2000a)). When heteroscedasticity is present in a standard probit model, inconsistent and inefficient estimators are produced. In spatial models, spatial autoregressive patterns frequently entail heteroskedasticity. Therefore, proper algorithms to calculate estimators are needed. The estimation part will be covered in section 4 of this document.

Another field of research is that which contemplates time within the spatial probit model. Until now, in the models already presented, it is assumed that everything happens in the same time frame. *Dynamic Spatial Ordered Probit* (DSOP) models assume that the autoregressive effect does not occur at the same moment in time, therefore the modeling should be dynamic. On this topic, Wang and Kockelman (2009) analyzed land use over the years.

The specification of this model is as follows

$$\begin{aligned} Y_t^* &= X_t\beta + \lambda Y_{t-1}^* + \theta_t + e \\ e &\sim N(0, \sigma^2 I) \end{aligned} \tag{1.28}$$

where λ in this case is the temporal autocorrelation parameter and θ is a vector that contains the spatial autoregressive parameter and therefore the neighborhood influence.

$$\theta_i = \rho \sum w_{ij}\theta_j \tag{1.29}$$

The last extension of the probit models are the *State transfers at different moments in time* proposed by Elhorst et al. (2013). This paper considers two spatially lagged variables because it takes time into consideration in the model. It is considered that those observations in which the observable dependent variable has become Y = 1, do not affect the observations in which the continuous dependent variable Y = 0. Therefore, the lagged spatial variable of those observations that are already Y = 1 can be considered exogenous.

The specification of this model is as follows

$$Y_t^{0*} = X_t^0\beta + \rho W_t^{00}Y_t^{0*} + \delta W_t^{01}S_t + e_t^0 \qquad e_t \sim N(0, \sigma^2 I) \tag{1.30}$$

where $W_t^{00}Y_t^{0*}$ is the endogenous effect for those observations closed to each other which remain in Y=0, then W_t^{01}\$S_t \$ which is the exogenous interaction effect for those observations closed to each other with observations that moved to Y=1.

The reality is that there are not many studies with this type of structure given the complexity of the subject or the availability of sufficient data. However, from a more logical point of view, it makes sense that, for example, the behaviors of economic agents influence each other at different points in time.

1.3 Testing Spatial dependence

One of the development points within spatial econometrics that generated a great deal of interest was the development of tests to detect autocorrelation in the residuals from classical models. Cliff and Ord (1973) demonstrate how the I-Moran statistic can be used to assess the autocorrelation of the residuals from a classical regression estimated by ordinary least squares. Obviously, this first investigation was carried out on a standard Gaussian regression model. A decade later, tests for maximum likelihood estimation began to be developed, specifically the Lagrange Multiplier test (LM test) (Anselin, 1988). The latter also for a standard Gaussian environment. Just before reaching the 21st century, robust LM tests were developed to test whether the residuals of a linear model come from a spatial SEM or SAR structure (Anselin et al., 1996). At the same time, and for the case that concerns us in this paper, tests were developed for those models in which the dependent variable is limited.

In the next subsections, the most used tests both in the academic field and in the main spatial econometric software, focused on binary variables, are briefly exposed. In the first part, tests that analyzes the residuals of a probit model to detect the presence of spatial dependence are presented. These tests are either global, affecting the entire map within the scope or local, affecting very specific areas of the map. Secondly, the statistics that work directly on the binary endogenous variable are exposed, determining the presence of spatial autocorrelation both in the global and local framework. Finally, the generalized algorithm to cluster search method is presented given a Bernoulli spatial process.

1.3.1 Tests on Standard Probit Model Residuals

Spatial correlation is the relationship of what happens at a point i in space and what happens in adjacent places. A random variable is spatially autocorrelated when the values observed in i depend on the values observed in contiguous areas. The standard linear probit model assumes independence between the observations and constant variance. When we have a spatial autoregressive process, this independence is breached and inefficient and inconsistent parameters are generated.

Probit Standard Model:

$$P(y_i = 1/x_i) = P(y_i^* > 0/x_i) = P(e_i < x_i\beta/x_i)$$

Spatial Probit Model SAR:

$$P(y_i = 1/x_i) = P(y_i^* > 0/x_i) = P(e_i < \frac{(I - \rho W_1)^{-1} x_i\beta}{\sigma_i}/x_i)$$

$$\Omega = E(ee') = \sigma^2((I - \rho W_1)^{-1})((I - \rho W_1)^{-1})')$$

Spatial Probit Model SEM:

$$P(y_i = 1/x_i) = P(y_i^* > 0/x_i) = P(e_i < \frac{x_i\beta}{\sigma_i}/x_i)$$

$$\Omega = E(ee') = \sigma^2((I - \lambda W_1)^{-1})((I - \lambda W_1)^{-1})')$$

Since the correct specification of our model is not known a priori, nor whether the data comes from a standard or spatial theoretical model. It is vitally necessary to have tests that at least diagnose whether the errors evaluated in the standard model are spatially correlated or not. Although, what these tests would not tell us is what kind of spatial specification we would be talking about.

The tests developed to date will be presented below. Following the notation of Amaral et al. (2013), in which the different autocorrelation tests for spatial probit for the SEM models are compared. In all these test

$$H_0 : \text{Disturbances are Spatially Independent}$$
$$H_1 : \text{Disturbances are Spatially Dependent of each other}$$

1.3.1.1 Pinske and Slade (1998)

The first test that emerged was that of Pinkse and Slade (1998). The authors follow the robust LM test methodology that was being developed for standard Gaussian models.

$$PS_{test} = \frac{(e^t We)^2}{tr(WW + W'W)} \sim X^2(1) \tag{1.31}$$

Given that the measurement of the error is not as obvious as in the case of standard models, throughout the academic literature different approaches have emerged as to what should be the correct measurement of the error in the probit model. In this case the authors choose the error in its most direct form corrected for its variance. Then,

$$e_i = \frac{y_i - \Phi(\hat{y}_i^*)}{\sqrt{(\Phi(\hat{y}_i^*)(1 - \Phi(\hat{y}_i^*)))}} \tag{1.32}$$

1.3.1.2 Pinske (1999 and 2004)

The second variation to the problem would be made years later in (Pinkse, 1999, and Pinkse (2004)) with a modification of the previous LM test that himself proposed.

$$P_{test} = \frac{(e^t We)^2}{\sigma^4 tr(WW + W^t W)} \sim X^2(1) \tag{1.33}$$

where,

$$\sigma^4 = (n^{-1} \sum_i^n \frac{\Phi'(\hat{y}_i^*)^2}{\Phi(\hat{y})(1 - \Phi(\hat{y}_i^*))})^2$$

and in this case the errors are written as

$$e_i = \frac{(y_i - \Phi(\hat{y}_i^*)(\Phi'(\hat{y}_i^*))}{\Phi(\hat{y})(1 - \Phi(\hat{y}_i^*))} \tag{1.34}$$

1.3.1.3 Generalized I-Moran (2001)

Kelejian and Prucha (2001a) generalizes de famous I-Moran test (Moran, 1950) in order to be able to apply it to those problems in which the dependent variable is binary.

$$KP_{test} = \frac{(e^t We)^2}{\sqrt{tr(W\Sigma W\Sigma + W^t \Sigma W\Sigma)}} \sim N(0, 1) \tag{1.35}$$

where

$$\Sigma = matrix(diag((\Phi(\hat{y}_i^*)(1 - \Phi(\hat{y}_i^*)))))$$

And in this case the error takes its simplest form

$$e_i = y_i - \Phi(\hat{y}_i^*) \tag{1.36}$$

1.3.2 Tests on Endogenous Variable

The tests that have been named so far are based on the analysis of the residuals of a standard linear model. Tests have also been developed that directly assess the dependent binary variable to determine its spatial autocorrelation. In this sub-section we will analyze some of the most common tests in this area. This type of test, in addition to providing information on the existence of global or local spatial dependence, also helps the researcher when deciding on those adjacency matrices that are theoretically known, however in practice they hide a very extensive background on which is the ideal W matrix that best represents the spatial dependence of the data.

1.3.2.1 Joint-Count statistics

Join-Count statistics (Cliff and Ord, 1973) are used to test global spatial autocorrelation pattern of both binomial and multinomial variables. Join-Count statistics tests the null of the random co-localized pattern by counting the number of each possible joins between neighbors. In the binomial context, possible joins are J11 (Black-Black), J00 (White-White) and J10 (Black-White). The statistics J11, J00, and J10 count the observed number of joins and compare them with the expected number under the null (J'11, J'00, and J'10). In order to join elements, the binary weight matrix W is needed to establish a connectivity criterion. In particular, the elements of W, wij (i,j=1,. . . ,n) have a value of 1 if elements i and j are neighbors and 0 otherwise. Two elements "i" and "j" are joined if the j-element belonged to the set of k-nearest i-element. From this connectivity criterion, the Join-Count statistics (J11, J00, and J10) are defined in the following equations:

$$\begin{aligned} J11 &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}11_{ij}; \\ J10 &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}10_{ij}; \\ J00 &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} - (J11 + J10) \end{aligned} \tag{1.37}$$

The statistics described are distributed as an asymptotic normal distribution. Spatial co-localized patterns which result from the application of Join-Count tests can be positive or negative. A positive co-localized pattern indicates a spatial structure in which there is a high probability of finding customers who belong to category 1 or 0, surrounded by customers who fall into the same category. A negative result reveals the spatial interconnection of customers who fall into different categories and when the spatial distribution is random, no spatial co-localized pattern can be ascertained.

1.3.2.2 Local Join-Count statistic

Local Join-Count statistic (Anselin and Li, 2019) is the local spatial correlation indicator for binary variables. Following (Anselin, 1995), the test is based on the idea that the global spatial correlation can be broken down into a number of local indicators. Unlike the global join count test, the local test only makes sense for the cases 11 (Black-Black) in which location "i" has an occurrence=1 (x_i=1) comparing to the number of occurrences=1 of neighbors around. The test would be significative when locations around "i" have more occurrences equal to one (x_i=1) than a simply random pattern.

$$LJ11_i = x_i \sum_{j=1}^{n} w_{ij} x_j; \tag{1.38}$$

where x_i is the binary value at location "i" and x_j is the binary value at location "j".

The problem with the local join count statistic is the methodology to assess the significatively of each point. Depending on the size of the data set and the inbalance of binary variable, the interpretation of local patterns should be treated carefully.
The procedure to calculate the probability of the existence of a local pattern, can be done following properties of hypergeometric distribution or with a more traditional approach which is based on the conditional permutation test proposed by Anselin in (Anselin, 1995) for the LISA statistic.

1.3.2.3 Q statistic

Ruiz et al. (2010) uses symbolic dynamics to undertake the Q test. The Q test basically determines both the strength and the significance of spatial associations. For a data set of N elements, k is the number of categories of the variable to study. In the case studied in this work, since it is a dichotomous variable k = 2. However, the authors establish the Q test for k> = 2 in case of multinomial data. m-1 is defined as the number of neighbors that each observation can have. X, therefore, is a vector of length m, which contains, in a sorted way, the value of the studied variable at each point in space and the value of the

closest m-1 observations. The symbols σ_i will be each unique combination of Xm. There will be at most a number of k^m symbols. The probability of occurrence of each symbol p_{σ_i} will be the frequency of each symbol over the total number of observations.The entropy function will be defined as

$$h(m) = -\sum_{\sigma} p_\sigma ln(p_\sigma) \tag{1.39}$$

In such a way that the entropy varies between 0 when there is no entropy and $ln(k^m)$ when the process is totally random. The more autocorrelation there is across the map, the less entropy will be detected with symbolic use. And vice versa, the more random the spatial process, the greater entropy there will be.

The Q test is defined as follows

$$Q(m) = 2(\frac{N-m}{m-r} + 1)[ln(k^m) - h(m)] \sim \chi^2_{k^m-1} \tag{1.40}$$

where r is the overlapping degree of X.

Based on this Q test, it is proposed in Páez et al. (2013) an indicator of the spatial fit of a model. When a probit model is built in which the endogenous variable has spatial autocorrelation, it is interesting to verify that after modeling the endogenous expected variable still contains this spatial pattern. The statistic constructed for the assessment is

$$\frac{\hat{Q}(m) - Q(m)}{\hat{Q}(m) + Q(m)} \tag{1.41}$$

Where Q(m) is the Q test on the endogenous variable and $\hat{Q}(m)$ is Q test on the expected endogenous variable. The null hypothesis is that the vector with the estimated value of Y contains the same level of spatial association as the real variable.

1.3.2.4 Scan statistic for binary data

The scan statistic is a very well-known method to detect spatial clusters in those areas where there is a significant increase of a certain event. The spatial clustering was firstly studied by Naus (1965) both in one and two dimensions. In Naus (1974), the problem is extended to cover the Bernoulli process and then this algorithm is generalized in Kulldorff (1997). The latter paper also allows for different types of cluster shapes and sizes which

makes the algorithm very attractive in those research where the area of the cluster is unknown. Following Kulldorff (1997) notation, the hypothesis tested are

$$H_0 : p = q$$
$$H_1 : p > q$$

where p and q are the probabilities for the occurrence of certain event inside and outside a window or potential cluster Z. This Z window is inside a total spatial scope under study called G. The idea is to maximize the likelihood function

$$L(Z) = p^{c_z}(1-p)^{n_z-c_z}q^{C-c_z}(1-q)^{N-n_z-(C-c_z)} \tag{1.42}$$

where n_z is the number of observations inside the window Z, c_z is the number of positive cases inside the window Z, C is the total number of cases in the scope G, N is the total number of observations in G. The expression also can be written as

$$L(Z) = (\frac{c_z}{n_z})^{c_z}(1-\frac{c_z}{n_z})^{n_z-c_z}(\frac{C-c_z}{N-n_z})^{C-c_z}(1-\frac{C-c_z}{N-n_z})^{N-n_z-(C-c_z)} \tag{1.43}$$

In order to make statistical inference, the likelihood ratio is calculated as follows

$$LR = \frac{L(Z)}{(\frac{C}{N})^C(\frac{N-C}{N})^{N-C}} \tag{1.44}$$

By iterating the calculation through the different points of the map the most likely cluster will be reached when the subset Z_i maximizes the LR equation.

All the tests presented help the analyst to make a decision on the type of model to apply to make a correct approach to reality. The construction of an econometric model should be independent of the a priori judgment of the researcher. The detection statistics of the presence of spatial autocorrelation objectify the entire process towards the correct specification. Not to mention the guidance in the search for focused and efficient estimators.

1.4 Probit Model Estimation

It was said at the beginning of the chapter that spatial probit models have not received as much attention as classic regression models. One of the reasons why this fact occurs

is due to the added complexity to solve these models and the high computational cost which increases as the data grows. Over the last few years, algorithms capable of dealing with the estimation of the parameters of these spatial probit models have emerged. The first line of research was to obtain efficient and consistent estimators. And the second line more focused on reducing the computing times necessary to solve the problem.

The algorithms that have been used to solve the problem were primarily the *Expectation-Maximization* (EM) proposed by McMillen (1992). In this paper, a problem of spatial autocorrelation and heteroskedasticity is exposed. The way to solve it will be through replacing the unobservable latent variable with the expected value and then iterating until convergence. The problem with this method, admitted by its own author, is the computational cost that it mainly involves. The method requires many iterations and sub-iterations which make the process very slow. Pinkse and Slade (1998) proposed the solution by using the *Generalized method of moments* (GMM), which provides estimates in a very efficient time, however the estimates are not very precise as shown in Calabrese and Elkink (2014), in a revision of the estimations of spatial probit models. LeSage (2000a) uses *GIBBS bayesian sampling* to estimate a heteroskedastic spatial autoregressive problem. As in the EM algorithm, the unobservable latent variable is replaced by the expected value. In this model, the conditional probability of each of the parameters is specified. Then, Beron and Vijverberg (2004a) proposed to use the *Recursive Importance Sampling* (RIS) based on MonteCarlo simulations to capture the true coefficients. According to Calabrese and Elkink (2014), GIBBS and RIS are very accurate in the estimates, however for samples greater than 1500 observations the computation time is very high. In Klier and McMillen (2008) a modification of GMM algorithm is proposed by linearizing it around a starting point where autoregressive parameters are set to zero (LGMM)[1]. In this case, the algorithm is prepared for logit instead of probit. Similar to the standard GMM model, the model produces bias estimators especially when there are not many observations or when the autoregressive parameters are high. Last but not least, Martinetti and Geniaux (2017) proposed directly to maximize the likelihood function by approximating it with an analytical method.

1.4.1 Estimating spatial probit models by maximum likelihood

Now, paying attention to *spatial autoregressive specification models* (SAR). These models appear when there is an endogenous spatial structure of dependency.

[1] Although this algorithm is focused on solving the problem using a logit, it has been included within the brief description given its relevance

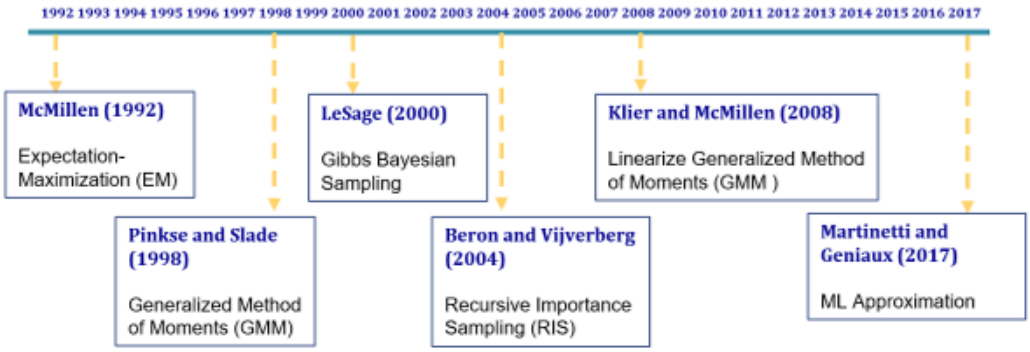


Figure 1.1: Timeline of Spatial Probit Estimation Techniques

$$Y^* = (I - \rho W_1)^{-1}(X\beta + u) \tag{1.45}$$

therefore the whole error term is

$$\psi = (I - \rho W_1)^{-1}u$$

and it's variance can be expressed as,

$$\Omega = E(\psi\psi') = \sigma^2((I - \rho W_1)^{-1})((I - \rho W_1)^{-1})') \tag{1.46}$$

As we are in a probit environment, then we assume a multivariate normal distribution for the likelihood function with a covariate structure.

$$L(\beta, \rho) = \frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \int_{a_1}^{b_1} \int_{a_2}^{b_2} ... \int_{a_n}^{b_n} e^{-\frac{t^t \Omega^{-1} t}{2}} dt \tag{1.47}$$

The intervals of integrations depend on the value of the observable Y.

$$\text{if } y_i = 1 \text{ then: } a = -\infty; b = ((I - \rho W_1)^{-1})X\beta$$
$$\text{if } y_i = 0 \text{ then: } a = ((I - \rho W_1)^{-1})X\beta; b = \infty$$

Now, let's examine *spatial error model* (SEM) which contains the spatial autoregressive parameter in the error structure.

$$Y^* = X\beta + (I - \lambda W_2)^{-1}e \tag{1.48}$$

now, the whole error term is

$$\psi = (I - \lambda W_2)^{-1}u$$

and it's variance can be expressed as

$$\Omega = E(\psi\psi') = \sigma^2((I - \lambda W_2)^{-1})((I - \lambda W_2)^{-1})^t) \tag{1.49}$$

By maximizing the likelihood, the standard and autoregressive coefficients could be obtained.

$$L(\beta, \rho) = \frac{1}{(2\pi)^{n/2}|\Omega|^{1/2}} \int_{a_1}^{b_1} \int_{a_2}^{b_2} ... \int_{a_n}^{b_n} e^{-\frac{t^t \Omega^{-1} t}{2}} dt \tag{1.50}$$

The intervals of integrations depend on the value of the observable Y.

$$\text{if } y_i = 1 \text{ then: } a = -\infty; b = X\beta$$
$$\text{if } y_i = 0 \text{ then: } a = X\beta; b = \infty$$

In case of combining both models, the *spatial autoregressive model with autoregressive disturbances* (SARAR) is obtained.

$$Y^* = (I - \rho W_1)^{-1}(X\beta + (I - \lambda W_2)^{-1}e) \tag{1.51}$$

Therefore the error term is

$$\psi = (I - \rho W_1)^{-1}(I - \lambda W_2)^{-1}u$$

and it's variance can be expressed as

$$\Omega = E(\psi\psi') = \sigma^2((I - \rho W_1)^{-1})(I - \lambda W_2)^{-1})((I - \rho W_1)^{-1}))^t((I - \lambda W_2)^{-1}))^t) \tag{1.52}$$

The N-Integrals for computing the multivariate normal distribution is equivalent to SAR and SEM case and the intervals for integrations

$$\begin{aligned} \text{if } y_i = 1 \text{ then: } a = -\infty; b = ((I - \rho W_1)^{-1})X\beta \\ \text{if } y_i = 0 \text{ then: } a = ((I - \rho W_1)^{-1})X\beta; b = \infty \end{aligned} \tag{1.53}$$

For each of the exposed models, the first derivatives would be taken with respect to the parameters associated with exogenous variables with respect to the spatial autoregressive parameters and they would be set equal to zero. Since there is no closed-formula for the n-integral, one has to resort to numerical methods for its approximation. According to (Wang et al., 2013) this task is an inordinate computational effort and unpractical from a pragmatic point of view. This is mainly the reason why until Martinetti and Geniaux (2017), only other type of options/methods had arisen instead of maximizing the likelihood of the problem.

1.4.1.1 ML Approx - Martinetti and Geniaux's (2017)

Martinetti and Geniaux (2017) proposed a conditional approximation method to obtain consistent and efficient estimations of the parameters. The idea the authors develop is based on the method of Mendell and Elston (1974b). The first part of the algorithm treats to reconstruct the multivariate normal probabilities as a product of univariate conditional probabilities.

$$\begin{aligned} P(t_1 \in (a_1, b_1)....t_n \in (a_n, b_n)) = \\ = P(t_1 \in (a_1, b_1)) \prod_{i=2}^{n} P(t_i \in (a_i, b_i)/(t_1 \in ((a_1, b_1).....t_{i-1} \in ((a_{i-1}, b_{i-1}))) \end{aligned} \tag{1.54}$$

In order to compute the approximation, the algorithm uses the Cholesky decomposition, so that,

$$t^t \Omega^{-1} t = t^t (C^{-1})^t C^{-1} t \tag{1.55}$$

After making the transformation t=Cz, then the the limits of the integral change

$$\begin{aligned} a'_n &= \frac{a_n - \sum_{i=n}^{n-1} C_{ni} z_{ni}}{C_{nn}} \\ b'_n &= \frac{b_n - \sum_{i=n}^{n-1} C_{ni} z_{ni}}{C_{nn}} \end{aligned} \tag{1.56}$$

which can be computed iteratively, while values of z are approximated.

$$z_i = \frac{\Phi'(a'_i) - \Phi'(b'_i)}{\Phi(b'_i) - \Phi(a'_i)} \tag{1.57}$$

$$\tag{1.58}$$

The result of the algorithm is

$$P(t_1 \in (a_1, b_1)....t_n \in (a_n, b_n)) \sim \prod_{i=1}^{n} \Phi(b'_i) - \Phi(a'_i) = \prod_{i=1}^{n} U_i \tag{1.59}$$

There are several computational aspects that make the process faster. The first is the rearrangement of the values. Instead of using the natural order of the observations, it is proposed to order descending based on the expected value. Although, as the authors cite, few rearrangements are advisable. Next, a set of numerical and computational techniques make the algorithm work much faster. The first is the use of sparse arrays since it frees up a lot of internal memory when performing operations. Another is related to the inverse of spatial multipliers. Instead of doing the inverse of large matrices, the Taylor expansion can be used as an approximation.

$$(I - X)^{-1} \sim I + X + X^2 + X^3 + X^4 + ...$$

The final part of Martinetti and Geniaux's algoritm estimates the coefficients. The authors present four different options to estimate the coefficients. The most complete set of options in terms of precision, although more expensive in time, is the one that estimates the coefficients based on the partial derivatives of the built-in likelihood function with respect to the parameters to be estimated using the gradient functions. The second set of options calculates the value of the autoregressive parameter using the likelihood with just this parameter to optimize and the rest of coefficients with the standard probit regression.

1.4.2 Alternatives to estimate spatial probit models by maximum likelihood

As stated in the introduction of this section, there were some other methods developed that solve the problem avoiding the maximization of the likelihood function. Next, the main methods that have been suggested will be briefly presented.

1.4.2.1 EM Estimation - McMillen (1992)

The method was proposed originally by Dempster et al. (1977a) as a method to maximize the likelihood of a model when missing data is present. McMillen (1992), uses this method to maximize the likelihood expression of spatial autoregressive models. The algorithm works with two steps, which are iterating until they converge and the model will have finished. The first step (E-Step) calculates the expected likelihood given some initial values. And the second step (M-step) maximizes the previous likelihood function.
Focusing on the probit case, in the E-Step the endogenous observable variable is replaced by the expected value of the latent variable. Therefore:

For SAR Model:

$$\begin{aligned} &\text{if } y_i = 1 \text{ then: } E(Y *_i | Y = 1) = (I - \rho W)^{-1} X\beta + \sigma \frac{\Phi'((I - \rho W)^{-1} X\beta)}{\Phi((I - \rho W)^{-1} X\beta)} \\ &\text{if } y_i = 0 \text{ then: } E(Y *_i | Y = 0) = (I - \rho W)^{-1} X\beta - \sigma \frac{\Phi'((I - \rho W)^{-1} X\beta)}{1 - \Phi((I - \rho W)^{-1} X\beta)} \end{aligned} \tag{1.60}$$

For SEM Model:

$$\begin{aligned} &\text{if } y_i = 1 \text{ then: } E(Y *_i | Y = 1) = X\beta + \sigma \frac{\Phi'(X\beta)}{\Phi(X\beta)} \\ &\text{if } y_i = 0 \text{ then: } E(Y *_i | Y = 0) = X\beta - \sigma \frac{\Phi'(X\beta)}{1 - \Phi(X\beta)} \end{aligned} \tag{1.61}$$

The next step (M-step) is to maximize the log-likelihood function to obtain new coefficientes.

For SAR Model:

$$K - (n/2) ln(\sigma_e^2) - (2\sigma_e^2)^{-1} ((I - \rho W_1)Y - X\beta)^t ((I - \rho W_1)Y - X\beta) + \sum_i^n (1 - \rho\omega_i) \tag{1.62}$$

For SEM Model:

$$K - (n/2)ln(\sigma_e^2) - (2\sigma_e^2)^{-1}(Y - X\beta)^t(I - \lambda W_2)^t(I - \lambda W_2)(Y - X\beta) + \sum_i^n (1 - \lambda\omega_i) \tag{1.63}$$

where ω_i are the eigenvalues of the contiguity matrix. This process is repeated until convergence.

The method obtains biased but consistent estimates; however the most important disadvantage of this algorithm is the computational time required to estimate the coefficients. Especially for samples larger than 1000 observations the computational cost is quite high. The tests carried out in the Monte Carlo exercise in the following section show, in addition to the mentioned high computational cost, how the algorithm tends to underestimate the spatial autoregressive parameter. These results are consistent with the results of Calabrese and Elkink (2014).

1.4.2.2 GMM Estimation - Pinkse and Slade (1998)

Pinkse and Slade (1998) developed a estimation method for Spatial Error Models (SEM) based on the Generalized Method of Moments (GMM). This method uses the error term $[e(parameters)]$ defined in 1.34. Obviously, the error depends on the parameters estimated β, ρ and λ. Therefore, the problem lies in minimizing the error based on the parameters.

$$EstimatedParameters = \arg\min e(parameters)'ZMZ'e(parameters) \tag{1.64}$$

where Z is an instrument matrix $Z = 1 + X + WX + W^2X + W^3X$ and M a symmetric positive matrix, although in Pinkse and Slade (1998) is equal to the identity matrix.

An extension of this method was proposed in Klier and McMillen (2008). The authors made a change in the function to be minimized by defining $M = (Z'Z)^{-1}$. There are five main steps in the iteration convergence. *First Step*: Take first approximated $parameters_0(\beta,\rho)$ initial estimators.*Second Step*: Calculate residuals (e_0) as in 1.36 and calculate the partial derivatives (G) of P over β and ρ. Where P is the logit model $P = exp(\hat{y}_i^*)/(1 + exp(\hat{y}_i^*))$. *Third Step*: Propose the regression G~Z and calculate $\hat{G}$. *Forth Step*: Calculate new estimation $parameters_1 = parameters_0 + (\hat{G}'\hat{G})^{-1}\hat{G}'e_0$. *Fifth Step*: Run until it converges.

GMM gets a good approach in ρ, but just when it is low[2]. The accuracy for the estimators of the independent variables in general is reasonable, but quite often it tends to calculate quite biased estimators. However, the most negative aspect of the algorithm is the time it takes to converge. When the sample size is greater than 1000, the computation time is likely to be greater than 20 minutes (see Monte Carlo Section). With respect to LGMM, it presents great accuracy, even improving GMM. Equivalently, when ρ is high, it fails to estimate it correctly, causing the model to misfit. However, its main benefit is the faster convergence speed.

1.4.2.3 GIBBS Estimation - LeSage (2000)

GIBBS sampling is a method based on Markov Chain Monte Carlo. This type of sampling makes sense when there are two or more dimensions involved in the problem. When sampling from original multivariate distribution is taught, a possible solution is GIBBS sampling which proposes a simpler and faster method by sampling from conditional distributions. In LeSage (2000a) this algorithm is proposed to solve the spatial probit regression problem. The first consideration made is the assumption that (Y) is observable. So the conditional distribution given an observable (Y) and known W is

$$p(\rho, \beta, \sigma/y, W) \propto |I - \rho W|\sigma^{-(n+1)} e^{\frac{-e'e}{2\sigma^2}} \tag{1.65}$$

By assuming that β and ρ are known, it is easy to see that $\sigma^2 \sim \chi^2_n$ to extract random draws to perform the sampling for this estimator. For β assuming the rest of the parameters known

$$\begin{gathered} p(\beta/y, W, \rho, , \sigma) = N(\tilde{\beta}, \sigma_e^2 (X'C'CX)^{-1}) \\ \text{for SAR: C=I and } \tilde{\beta} = (X'X)^{-1}(X'(I - \rho W)Y) \\ \text{for SEM: C=}(I - \rho W) \text{ and } \tilde{\beta} = (X'(I - \rho W)'(I - \rho W)X)^{-1}(X'(I - \rho W)'(I - \rho W)Y) \end{gathered} \tag{1.66}$$

Finally, the random draws for ρ are taken from expression 1.65 above. In order to faster calculations, not all the draws are taken, the "Metropolis Sampling" is used (Metropolis et al., 1953). Additionally, just the estimates consistent with the restriction $1/\min(\text{eigenvalue\_of\_W}) < \rho < 1/\max(\text{eigenvalue\_of\_W})$ (Anselin, 1988) are taken into account.

[2]The Monte Carlo experiment performed show that when ρ is higher than 0.5, the algorithm underestimates it.

This method achieves a very good precision both in the adjustment of the exogenous factors and in the autoregressive parameter. It must be said that the way the algorithm is proposed through MCMC sampling, as indicated by Calabrese and Elkink (2014) and Martinetti and Geniaux (2017), it could take a long time to converge. However, the computational development carried out by Wilhelm and de Matos (2013) in R, makes the algorithm much faster, achieving proper estimates in a few minutes when the sample is less than 10 thousand records.

1.4.2.4 RIS Sampling - Beron and Vijverberg (2004)

In this case another technique of sampling is used to evaluate the n-dimensional normal probability. Beron and Vijverberg (2004a) propose the *Recursive Importance Sampling* to solve the spatial probit problem. RIS sampling is a method to simulate discrete choice probabilities in a multivariate probit model. This algorithm is basically the same as GHK Monte Carlo sampling.

Let's create the diagonal matrix Z where the elements of the diagonal take the value of 1 if y_i=0 and the value of -1 if y_i=1, so that:

$$diag(Z) = 1 - 2y_i \tag{1.67}$$

The variance-covariance matrix for ν=Ze would be Ω_ν= Z Ω Z'. Where ω is defined in (1.46,1.49,1.52). Therefore ν will be distributed as N(0,Ω_ν). Ω_ν^{-1} can be rewritten using Cholesky as Ω_ν^{-1}=A'A and let η=Aν. So, η is i.i.d standard normal. A^{-1} will be an upper triangular matrix whose principal diagonal is always positive B=A^{-1}, thus, Bη=ν. The upper limit of ν will be V=$-Z(I-\rho W)^{-1}X\beta$. Then, the recursive iterative process to define η is

$$\begin{aligned} \eta_n &= \frac{V_n}{b_{nn}} \leftarrow \eta_n, 0 \\ \eta_j &= \frac{V_j - \sum_{i=j+1}^{n} b_{ji}\eta_i}{b_{jj}} \leftarrow \eta_j, 0 \end{aligned} \tag{1.68}$$

Then, to evaluate $\Pr(\nu < \mathrm{V})$

$$
\begin{aligned}
Pr(\nu < V) &= \int_{-\inf}^{V} \Phi'_n(\nu, 0, \Omega_\nu) = \\
&= \int_{-\inf}^{\eta_n} ... \int_{-\inf}^{\eta_1} \prod_{1}^{n} \Phi'_1(\eta_j) d\eta_1 ... d\eta_n = \\
&= \int_{-\inf}^{\eta_n} \frac{\Phi'(\eta_n)}{g^c(\eta_n)} ... (\int_{-\inf}^{\eta_2} \frac{\Phi'(\eta_2)}{g^c(\eta_2)} \Phi(\eta_1) g^c(\eta_2) d\eta_2) ...) g^c(\eta_n) d\eta_n
\end{aligned}
\tag{1.69}
$$

Where $g^c(\eta_j)$=$g(\eta_j)/G(\eta_j)$.Being g the density function and G the cumulative distribution function.

By generating a large number of random vectors of η and fulfilling $\eta_j \leq \eta_{j0}$. Given η_{n0} the rest of η can be calculated using 1.68. The final simulated parameters can be obtained doing

$$
\begin{aligned}
S &= NumberOfSimulations \\
EstimatedParameters &= \frac{\sum_{r=1}^{S} (\prod_{j=1}^{n} \Phi(\eta_{j,0,r}))}{S}
\end{aligned}
\tag{1.70}
$$

This method achieves good results of accuracy of the parameters. When ρ is high it tends to underestimate it slightly. The main and clearest disadvantage of the algorithm is the computation time. Without a doubt, it is the algorithm that we are presenting in this work that takes the longest to converge. The authors themselves at Beron and Vijverberg (2004a) carry out a Monte Carlo exercise in which they cannot use a large sample size due to computational limitations. With an i7 computer with 8GB of RAM, the algorithm could take an hour and a half to converge, which makes it unfeasible to use in practical terms.

1.5 Marginal Effects in Probit Models

Interpreting the way in which changes in the explanatory variables impact on the probability of occurrence (Y=1) is easy for the classical probit models while requires more care in the case of the autoregressive probit models. The reason is because of the spatial lag of the latent dependent variable WY* of the SAR probit model, changes in the value of the variable for observation j, influence observation i's probability. That is, now, the changes to the probability of Y=1 in location i are twofold: i) that induced by a change in the own-value of the variable, 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 observation, 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 observations different from 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).

1.6 R packages to estimate Probit Models

Over the last few years the popularity of spatial econometrics has grown. This has meant that specific packages or software have been created to solve problems on the matter. The first softwares with spatial econometric elements were SpaceStat developed by Luc Anselin in 1992 and a little later Spatialstats developed by Stephen Kaluzny in 1996. Later, thanks to the contributions of James LeSage or Paul Elhorst, specific modules were created in Matlab, which In the first decade of the 21st century, it gave a lot of visibility to this type of methodologies. Since then, both specific free software such as GeoDa (Anselin, 2003) or commercial software such as ArcGIS (Scott and Janikas, 2010) or Stata (StataCorp, 2017) with toolboxes on spatial econometrics have emerged. Undoubtedly, one of the languages that more researchers in spatial statistics have adopted in recent years and that has received the most contributions has been R and according to Bivand et al. (2021), it contains by far the richest variety of options.

The reference packages for purely geographic themes are *sp* and *raster*. In them the user can access the creation of polygons, spatial structures or manipulation and writing of different types of spatial objects. In fact, as can be seen in Bivand et al. (2008) (page 5), practically all the spatial packages developed to date depend on or directly import the *sp* library. Focusing on more analytical aspects, the *spdep* library is one of the first and great contributions that Roger Bivand has and many co-authors review and incorporate new features into this package. Its main functions are aimed at creating adjacency matrices with polygons or points, it contains the main global and local autocorrelation tests and various display functions. The best known library for spatial regression models is *spatialreg*, in which we find the estimation functions for specifications of the models that have been named in section 2. Although, the developments have been made only in this package for the classical Gaussian environment.

The reality is that there is not much variety of packages in terms of models where the dependent variable is limited like probit models. Several of the packages that exist are linked to the methodological development provided in the academic literature and which in turn makes the contribution to free software. The package *McSpatial* developed by Daniel McMillen in McMillen (2013), contains the functions to solve the spatial probit

and logit models using the Generalized Method of Moments(GMM), its linearized extension (LGMM) and also Maximum Likelihood (ML). The next package to appear is the *spatialprobit* developed by Stefan Wilhelm and Miguel Godinho de Matos in Wilhelm and de Matos (2015), in which the Bayesian estimation of the coefficients for probit and tobit by using MCMC and GIBBS sampling. As previously reported, the use of sampling techniques is quite efficient but extremely slow for large data samples. To mitigate these computational restrictions, the package uses sparce arrays, compiles the code in fortran, and uses parallelization. Another advantage of the package is that it allows to obtain the direct average effects, the total effects and indirect effects as difference of the two. The library calculates marginal effects quite fast by using QR-decomposition instead of inverting the n x n matrix $(I - \rho W)$ in each MCMC loop. This process is explained in great detail in Wilhelm and de Matos (2013). Another Bayesian method for solving spatial problems is based on the Integrated Nested Laplace Approximation (INLA). INLA is implemented in the *R-INLA* package providing posterior marginals at quite efficient times. Gómez-Rubio et al. (2021) has presented very recently the benefits and accuracy of these methods for dichotomous spatial processes. Finally in our review, the package *ProbitSpatial* which uses the method of Martinetti and Geniaux (2017) and it was programmed by the authors as well. They propose to solve the spatial probit model by maximum likelihood using approximation methods inspired by Mendell and Elston (1974b). This last package presented, as far as we know, is the only one that includes the SARAR model as an option to estimate its parameters. It allows the user to estimate the coefficients with conditional and unconditional expectations, as well as, tpo change the level of precission in order to minimize the time to converge. Additionally, in order to speed up calculations, the code is programmed in R and C ++ using the Rcpp and RcppEigen libraries. It also contains a function to show the marginal effects.

Without a doubt, this list of packages and features will continue to grow for years to come. After three decades of contributions, it can be said that there is a robust scientific community developing novel advances in spatial econometrics. Both in terms of computational efficiency and new algorithms to give light on the vastness of georeferenced data that is being captured today.

1.6.0.1 Comparisons between Estimations

Once the entire universe of the spatial probit has been presented, the objective of this subsection is to perform an analysis of the estimation of the different algorithms in R. Throughout a simple Monte Carlo experiment, both the bias in the estimated coefficients and the execution time will be tested. Both the bias of the estimator and the time consumed depend on the sample size and the degree of spatial dependence, so we will vary these variables to see their impact.
The Monte Carlo experiment is to be performed on the specification of a SAR

model (see Section 1.17). In each execution an independent variable distributed as a Normal(μ=1,σ=2) is generated and a regular lattice of $\sqrt{n} \times \sqrt{n}$ is also created randomly. The observations will be randomly distributed on these grids. The adjacency criterion of the observations will be the "Rook" criterion and then the matrix will be standardized. In order to create the vector of simulated Y^*, the residuals will have a distribution of a Normal(0,1), so we use ideal conditions to see the performance of these algorithms. The final Y^* will have the theorical beta parameters β=(1, -0.5). As said, in order to test the algorithms different sample sizes are proposed n=(100,400,900,1600,2500). When the sample size makes the algorithm not converge or the amount of time required is disproportionate[3], the Monte Carlo iteration stops and goes to the next iteration. The ρ parameter is chosen between ρ=(0.3,0.5,0.7), to be able to see the influence that the spatial dependency parameter has on the properties of the estimation algorithm. All this study is undertaken with an Intel(R) Core(TM) i7-6500U CP @ 2.50GHz 2.59GHz with 8GB of RAM with no parallel simulations.

$$\begin{aligned} &\text{Model Simulated:} \\ Y^* &= 1\beta_1 + X\beta_2 + \rho W Y^* + u \\ u &\sim N(0,1) \end{aligned} \tag{1.71}$$

Before going into all the details of the experiment, it must be said that the fastest[4] and most imprecise way to model a phenomenon with a spatial delay of the dependent variable is using the standard Probit. Bias for different ρ and sample size are contained in Table 1.1. The function used to calculate standard probits is the glm core function in R with family binomial and link "probit". As expected, estimators are biased and inconsistent, especially when ρ is high.

[3]Duration longer than 180 minutes

[4]The average for all the simulations presented in the analysis took less than a second to converge

Table 1.1: Bias in β_1 and β_2 using Standard Probit Algorithm

	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$
n = 100			
Bias β_1	0.2097	0.4025	0.6623
Bias β_2	-0.0064	0.0014	0.0588
n = 400			
Bias β_1	0.1936	0.3382	0.5754
Bias β_2	-0.0115	0.0355	0.0936
n = 900			
Bias β_1	0.1872	0.3329	0.5763
Bias β_2	0.0101	0.034	0.0792
n=1600			
Bias β_1	0.1549	0.343	0.5973
Bias β_2	0.0184	0.0371	0.0825
n=2500			
Bias β_1	0.1651	0.3313	0.5806
Bias β_2	0.0145	0.0389	0.094

The algorithms presented above designed to estimate the spatial probit coefficients will be tested. The algorithms that are optimized in public R functions are GMM, LGMM in *McSpatial* package, GIBBS in *spatialprobit* package and ML in *ProbitSpatial* package. Regarding EM and RIS, the implementation of Calabrese and Elkink (2014) for their Monte Carlo design is taken. The hyperparameters necessary to launch the algorithms are taken following the indications of the packages or papers that support them. GIBBS sampling have been run with 1000 MCMC iterations and with 100 burn-in discard. RIS sampling has been run with the same 1000 iterations based on Beron and Vijverberg (2004a) where a Monte Carlo study is performed with 1000 and 2000 simulations[5]. The simulations necessary to reach the optimum in the EM algorithm are also 1000 following indications from Calabrese and Elkink (2014) In the case of GMM and LGMM, we take the default convergence criterion established in the package programmed by Daniel McMillen. Finally, for the execution of the *ProbitSpatial* package we follow the suggestions in Martinetti and Geniaux (2017). By taking the conditional optimization method by which, it solves by standard probit methodology and then conditions on ρ. For each of the combinations of sample size, ρ and estimation algorithm, 500 simulations have been carried out.

1.6.0.1.1 Results under ideal conditions

The results of the bias of the estimator β and ρ are found in the Table 1.3 and Table 1.2 with the average of the iterations grouped by sample size and level of autocorrelation parameter and in Figure 1.4 and a box plot in Figure 1.2 with the distribution for each

[5]RIS sampling works reasonably well with 1000 iterations and obviously reduces computation times

algorithm solving spatial probit. The algorithms have been arranged in the chronological order in which their respective papers were published. EM in 1992, GMM in 1998, GIBBS in 2000, RIS in 2004, LGMM in 2008 and ML in 2017.

The first relevant aspect in the tables is the high volatility of the estimator when the sample size is equal to 100. Given such sample size, when ρ is small, we observe that the algorithms capture properly the estimator of β_1 and β_2. For such sample size and dependency factor, the algorithms that provide the best spatial dependence estimation are GMM, LGMM and ML. As we increase the spatial dependency parameter to 0.5 for n=100, the algorithms achieve a similar bias except GIBBS and EM, which fail to estimate the ρ parameter. All the algorithms tend to underestimate ρ except for LGMM which slightly overestimate it. When the spatial dependence is high ρ=0.7, basically the bias that we observe in more moderate ρ tends to be accentuated. Given the theorical parameters chosen or any other aspect of the Monte Carlo experiment, the reality is that EM algorithm achieves poor parameter fitting both for ρ, β_1 and β_2.

When the sample size varies between 400 and 900 observations, LGMM and RIS stop providing good fits when ρ=0.7. ML stands out as the algorithm that provides most centered estimations of the parameters. Especially when the sample size is increasing and when the spatial autocorrelation is moderate. Finally, for larger sample sizes such as n=1600, it can be seen that ML and GIBBS obtain centered results as a whole for ρ, $\beta_1$and β_2. For such sample size, the volatility of the estimators is lower than in the other sample sizes. LGMM continue overestimating the value of the spatial dependence specially when dependency is high and as a result, it generates bias in the independent variable beta estimator. RIS produces centered estimates in β_2 but underestimates ρ which is compensated overestimating the intercept (β_1). Finally, GMM produces very accurate estimates on ρ although present bias in β_2 and very volatile estimated β_1.

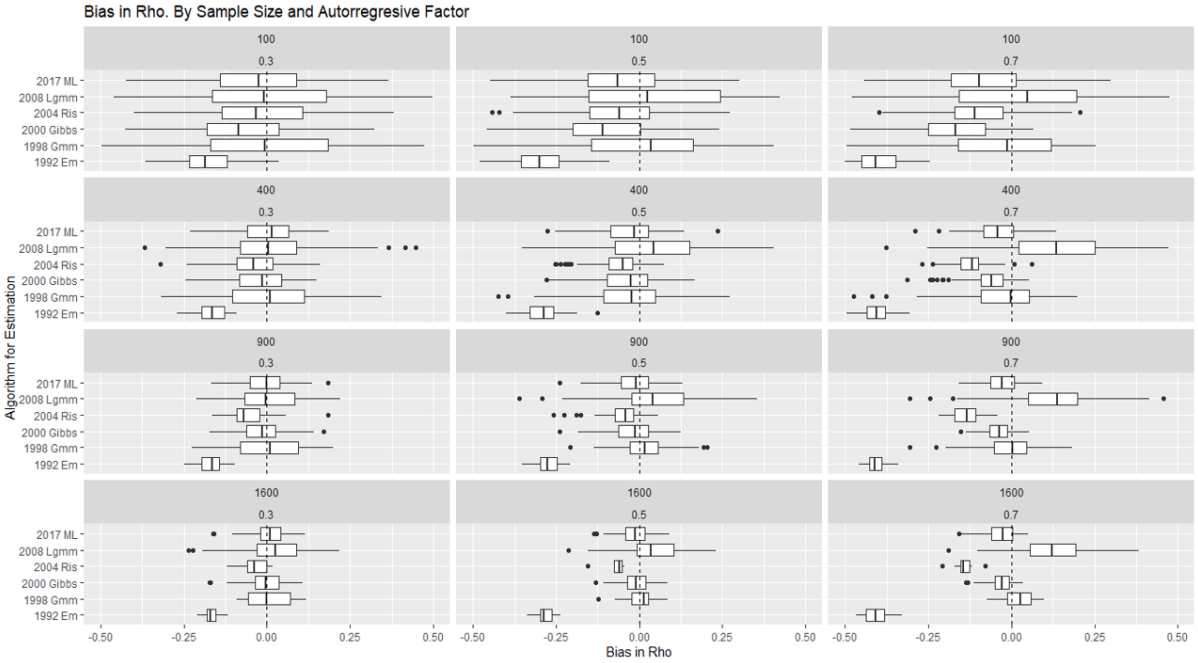


Figure 1.2: Bias in the ρ estimation under ideal conditions

Table 1.2: Bias in the ρ estimation under ideal conditions

		$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$
n = 100	1992 EM	-0.1753	-0.3032	-0.4521
	1998 GMM	-0.0325	-0.0483	-0.1808
	2000 GIBBS	-0.0751	-0.1221	-0.1982
	2004 RIS	-0.0284	-0.0700	-0.1090
	2008 LGMM	-0.0266	0.0423	0.0173
	2017 ML	-0.0341	-0.0726	-0.1105
n = 400	1992 EM	-0.1700	-0.2936	-0.4145
	1998 GMM	0.0042	-0.0229	-0.0212
	2000 GIBBS	-0.0268	-0.0404	-0.0625
	2004 RIS	-0.0415	-0.0621	-0.1191
	2008 LGMM	0.0155	0.036	0.134
	2017 ML	-0.0042	-0.0291	-0.0391
n = 900	1992 EM	-0.1705	-0.2752	-0.4079
	1998 GMM	0.0080	0.0153	-0.0130
	2000 GIBBS	-0.0157	-0.0173	-0.0392
	2004 RIS	-0.0538	-0.0579	-0.1365
	2008 LGMM	0.0066	0.0446	0.1249
	2017 ML	-0.0037	-0.0144	-0.0315
n = 1600	1992 EM	-0.1666	-0.2836	-0.4092
	1998 GMM	0.0083	-0.0051	0.0192
	2000 GIBBS	-0.0027	-0.0104	-0.0301
	2004 RIS	-0.0360	-0.0702	-0.1423
	2008 LGMM	0.0225	0.0432	0.1264
	2017 ML	0.0078	-0.0132	-0.0286

Table 1.3: Bias in β_1 and β_2 under ideal conditions

		$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$
n = 100	1992 EM β_1	0.1148	0.1908	0.4003
	1992 EM β_2	-0.0001	0.0218	0.0739
	1998 GMM β_1	0.2854	0.5021	1.3431
	1998 GMM β_2	-0.137	-0.2332	-0.3903
	2000 GIBBS β_1	0.1008	0.1481	0.269
	2000 GIBBS β_2	-0.0328	-0.0413	-0.011
	2004 RIS β_1	0.056	0.1089	0.24
	2004 RIS β_2	-0.0244	-0.0413	-0.0631
	2008 LGMM β_1	0.0215	-0.1044	-0.1059
	2008 LGMM β_2	-0.0094	0.0167	0.0446
	2017 ML β_1	0.0715	0.1325	0.3169
	2017 ML β_2	-0.0262	-0.0475	-0.0749
n = 400	1992 EM β_1	0.0926	0.1554	0.2531
	1992 EM β_2	-0.0002	0.0473	0.1036
	1998 GMM β_1	0.0229	0.0393	0.399
	1998 GMM β_2	-0.0244	-0.0051	-0.2085
	2000 GIBBS β_1	0.0335	0.0419	0.0573
	2000 GIBBS β_2	-0.0112	-0.0043	0.0106
	2004 RIS β_1	0.0462	0.0412	0.1328
	2004 RIS β_2	-0.0178	0.0127	0.0041
	2008 LGMM β_1	-0.0297	-0.1049	-0.3355
	2008 LGMM β_2	0.0089	0.0243	0.0706
	2017 ML β_1	0.0207	0.0429	0.1125
	2017 ML β_2	-0.011	-0.0056	-0.0122
n = 900	1992 EM β_1	0.097	0.1396	0.2438
	1992 EM β_2	0.0127	0.0432	0.0968
	1998 GMM β_1	0.0041	0.0134	0.0497
	1998 GMM β_2	0.003	-0.0172	-0.0307
	2000 GIBBS β_1	0.0138	0.0192	0.028
	2000 GIBBS β_2	-0.0006	-0.0041	0.0098
	2004 RIS β_1	0.0527	0.0373	0.1186
	2004 RIS β_2	0	-0.0049	0.0123
	2008 LGMM β_1	-0.0383	-0.1226	-0.3345
	2008 LGMM β_2	0.0124	0.0279	0.0743
	2017 ML β_1	0.008	0.0267	0.0903
	2017 ML β_2	-0.0006	-0.0057	-0.0063
	n = 1600			
n = 1600	1992 EM β_1	0.0647	0.1575	0.2617
	1992 EM β_2	0.0199	0.0447	0.1003
	1998 GMM β_1	0.0104	0.047	0.0193
	1998 GMM β_2	-0.0117	-0.0164	-0.0264
	2000 GIBBS β_1	0.002	0.0148	0.016
	2000 GIBBS β_2	0	-0.003	0.0121
	2004 RIS β_1	0.0219	0.0524	0.1117
	2004 RIS β_2	-0.017	0.0044	0.0151
	2008 LGMM β_1	-0.044	-0.121	-0.3388
	2008 LGMM β_2	0.0088	0.0291	0.0754
	2017 ML β_1	-0.0036	-0.0227	-0.0828
	2017 ML β_2	0.0001	-0.0027	-0.0028

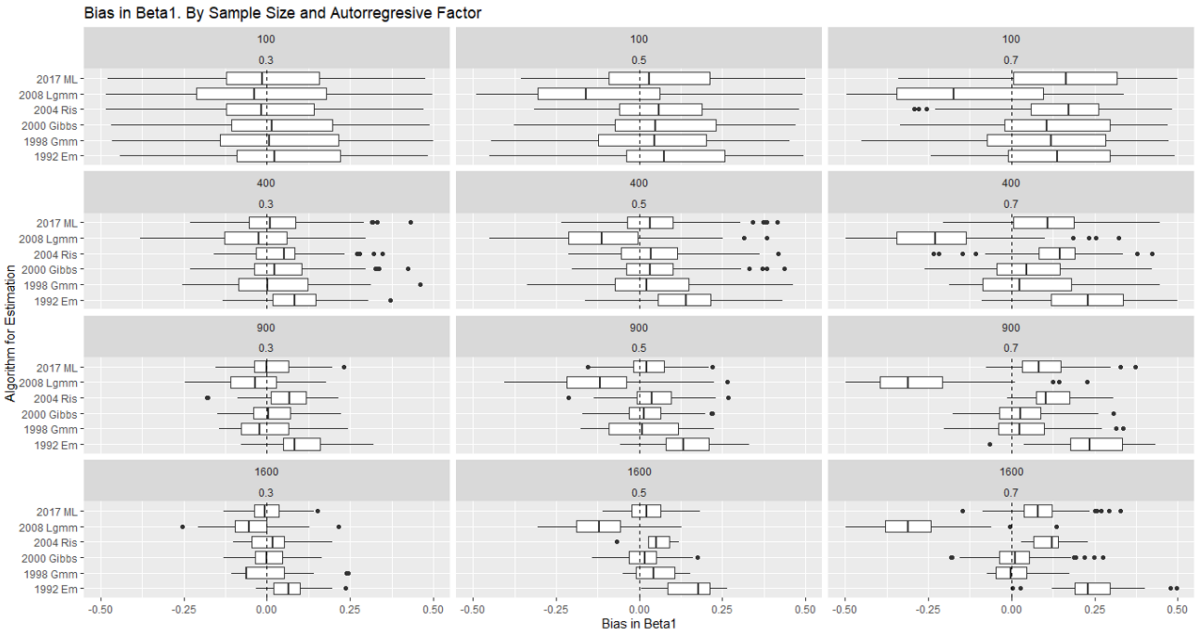


Figure 1.3: Bias in β_1 under ideal conditions

The results of the computational time taken by the executions of the study are found in Table 1.4 with the mean of the runs and in Figure 1.5 with the box plot with its distribution. The larger the actual spatial dependency in the data, the longer the computational time is generally required for the algorithm to converge. This increase is especially relevant in the Ris, GMM and EM algorithms. For these algorithms, the estimation becomes especially long from databases of more than a thousand records, obtaining durations greater than 20 minutes. However, the ML, LGMM and GIBBS algorithms, which are the ones that also have implementation and optimization in R, have quite affordable duration which makes them truly attractive for their use.

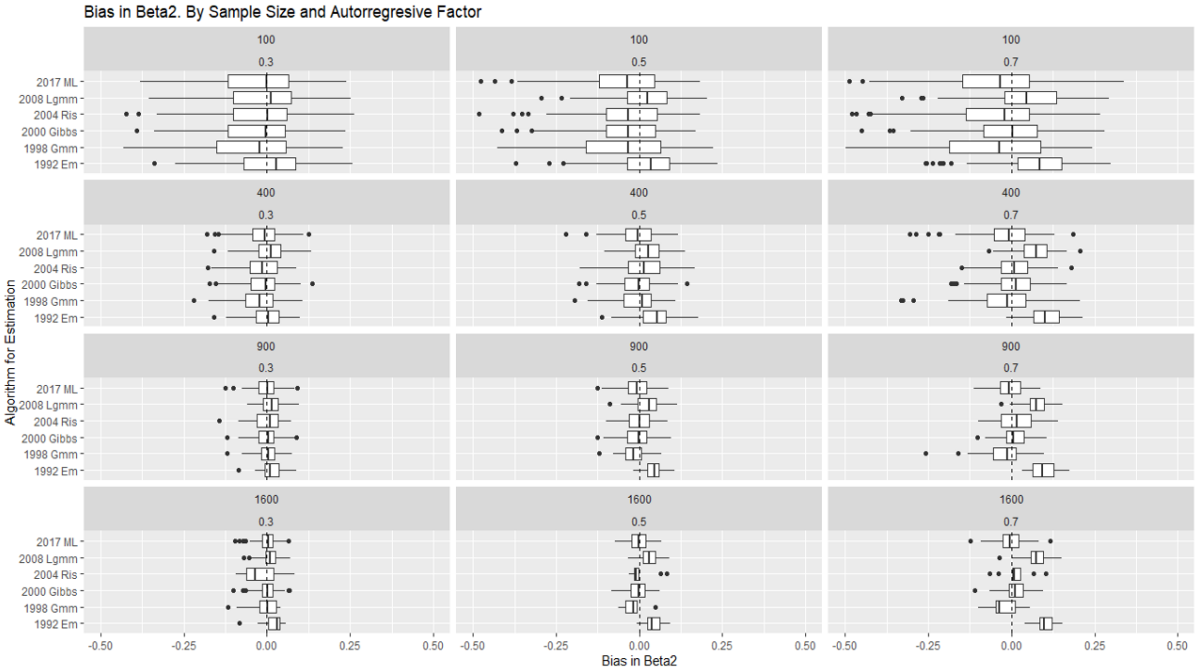


Figure 1.4: Bias in β_2 under ideal conditions

Table 1.4: Average Time in running spatial probit models under ideal conditions

	ρ=0.3	ρ=0.5	ρ=0.7
n = 100			
1992 EM	0.0082	0.0078	0.0112
1998 GMM	0.0118	0.0138	0.0181
2000 GIBBS	0.0527	0.0532	0.0524
2004 RIS	0.0677	0.1456	0.2305
2008 LGMM	0.0009	0.0002	0.0002
2017 ML	0.0008	0.0008	0.0009
n = 400			
1992 EM	0.1938	0.2202	0.2925
1998 GMM	0.5332	0.5521	0.7468
2000 GIBBS	0.0909	0.0912	0.091
2004 RIS	0.8298	2.1108	4.0725
2008 LGMM	0.0003	0.0003	0.0003
2017 ML	0.0043	0.0051	0.0054
n = 900			
1992 EM	2.016	2.5239	2.9057
1998 GMM	7.054	7.3515	9.8168
2000 GIBBS	0.1754	0.1772	0.1788
2004 RIS	3.6207	7.5709	19.0616
2008 LGMM	0.0007	0.0007	0.0006
2017 ML	0.0173	0.0195	0.022
n = 1600			
1992 EM	10.0955	13.3222	16.9005
1998 GMM	43.8724	45.2712	62.1566
2000 GIBBS	0.2924	0.2931	0.2965
2004 RIS	16.7811	19.975	26.003
2008 LGMM	0.0016	0.0012	0.001
2017 ML	0.0524	0.0559	0.0594

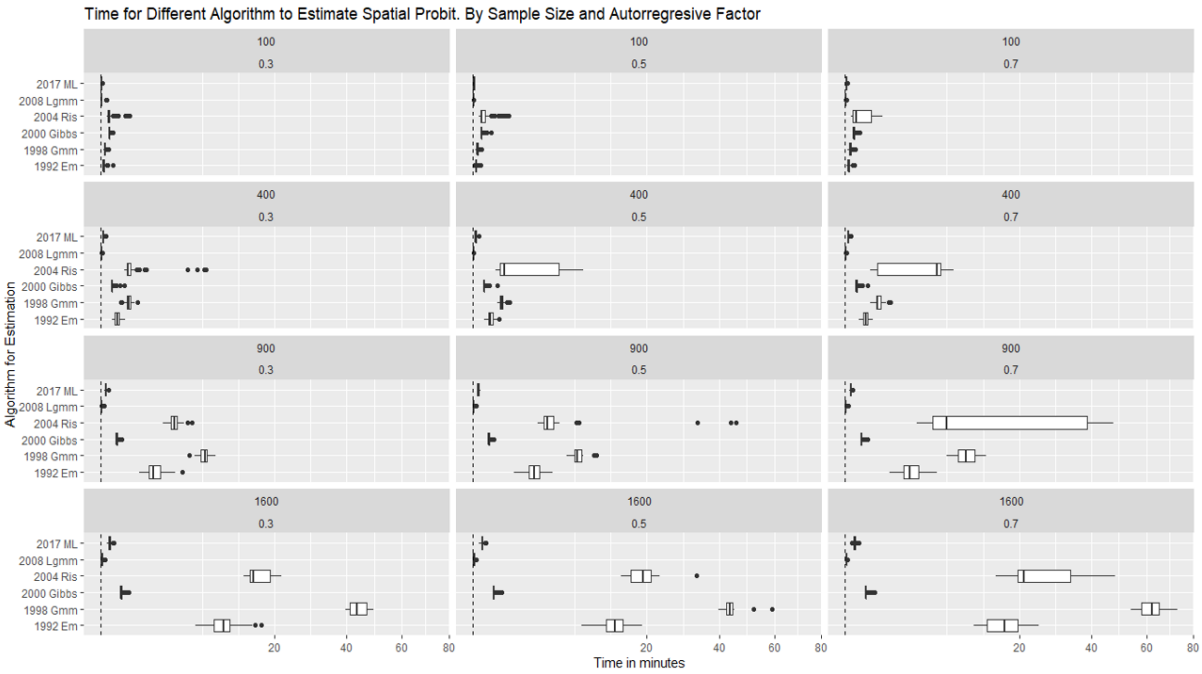


Figure 1.5: Box plot with Time (in Minutes) in running spatial probit models under ideal conditions

As the sample size increases, both ML, LGMM and GIBBS continue to show great performance. The purpose of these new data contained in the table 1.5 is to see the effects of notably increasing the sample size and also including the accuracy of each of the models through the Area Under the Curve ROC. LGMM (R spprobit function) is without a doubt the fastest in converging with the function's default hyperparameters. However, as it was the case with smaller sample sizes, it is the algorithm that provides the least centered estimators, especially when the spatial lag is high. This inaccuracy in the estimation of parameters is reflected in the ROC achieved by the algorithm, which is certainly low for ρ=0.7. GIBBS sampling (R SAR_probit_mcmc function) and ML (R ProbitSpatialFit function) provide fairly similar unbiased estimators achieving identical levels of accuracy. It could be said that the difference between both methods is the time to converge, being ML algorithm much more efficient, although to be fair, GIBBS sampling offers the possibility to change the number of draws to estimate the probability, so by changing number of iteration it could be possible to decrease the computational cost.

Table 1.5: Estimation of Bias and Computational time increasing the number of observations (ML, LGMM and GIBBS)

		$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$
		n=2500		
2000 GIBBS	Bias β_1	0.008	0.0177	0.0087
	Bias β_2	-0.0012	-0.0049	0.0102
	Bias ρ	-0.0046	-0.0096	-0.0247
	Minutes	0.7164	0.7016	0.711
	AUC	0.8397	0.8466	0.8558
2008 LGMM	Bias β_1	-0.0346	-0.1156	-0.3507
	Bias β_2	0.0088	0.0254	0.0741
	Bias ρ	0.0138	0.0449	0.1381
	Minutes	0.2192	0.2152	0.2166
	AUC	0.8398	0.8464	0.8358
2017 ML	Bias β_1	0.0025	0.029	0.0774
	Bias β_2	-0.0015	-0.007	-0.0037
	Bias ρ	0.0064	-0.0121	-0.0262
	Minutes	0.3356	0.3329	0.3455
	AUC	0.8397	0.8466	0.8558
		n=6400		
2000 GIBBS	Bias β_1	0.0021	0.0078	0.0038
	Bias β_2	-0.001	0.0005	0.0116
	Bias ρ	-0.0032	-0.0069	-0.0207
	Minutes	5.1863	5.1742	5.4498
	AUC	0.8403	0.8442	0.8559
2008 LGMM	Bias β_1	-0.0307	-0.116	-0.3525
	Bias β_2	0.0088	0.0311	0.0757
	Bias ρ	0.0014	0.0345	0.1349
	Minutes	3.2847	3.3043	3.4166
	AUC	0.8403	0.844	0.8473
2017 ML	Bias β_1	-0.003	0.0182	0.0732
	Bias β_2	-0.0006	0.0006	-0.001
	Bias ρ	0.005	-0.0127	-0.0237
	Minutes	3.9834	4.055	4.2414
	AUC	0.8403	0.8442	0.8559

1.6.0.1.2 Results under non-ideal conditions

In real data, we always find data with measurement errors, lack of information, endogeneity in the model and many other problems that make parameter estimation more difficult. In Table 1.6 we propose the assessment of bias, time and accuracy of the models under non-ideal conditions. The proposed model contains endogeneity and a spatially significant lagged variable X for the estimation of Y^* but not included in the estimation. Such that:

$$\begin{aligned} &\text{Model Simulated:} \\ Y^* &= 1\beta_1 + X\beta_2 + \rho WY^* + u \\ u &= -0.15X + 0.2WZ + e \\ &\text{Where:} \\ Z &\sim N(1,2) \text{ and } e \sim N(0,1) \end{aligned} \tag{1.72}$$

There are no significant differences among the accuracy achieved by the models. Given the simulated residual structure, the general precision tends to increase when the real ρ parameter is higher. LGMM presents a significant bias in the estimation of ρ regardless of the sample size. When the spatial dependence is moderate, the algorithm correctly adjusts the coefficients. However, just like under ideal conditions, when ρ gets large it introduces bias problems. As a result, LGMM shows a significant precision deficit. The AUC is almost identical for any sample size and ρ when estimating with GIBBS or with ML. Computational times increase by thirty percent when estimating with GIBBS for n=6400. And the bias in β and ρ is very similar for ρ greater than 0.5. For lower ρ, the bias presents greater volatility, making it difficult to choose between both models.

Table 1.6: Bias, Computational time in minutes and ROC under No Ideal conditions (ML, LGMM and GIBBS)

		$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$
	n = 2500			
2000 GIBBS	Bias β_1	0.1523	0.1488	0.1365
	Bias β_2	-0.0513	-0.0466	-0.0173
	Bias ρ	0.0014	-0.0008	-0.0302
	Minutes	0.6967	0.691	0.6934
	AUC	0.8568	0.8619	0.8751
2008 LGMM	Bias β_1	0.1028	0.0055	-0.2999
	Bias β_2	-0.038	-0.008	0.0501
	Bias ρ	0.0134	0.0317	0.1483
	Minutes	0.2168	0.2125	0.2116
	AUC	0.8568	0.8618	0.8546
2017 ML	Bias β_1	0.1465	0.1592	0.2247
	Bias β_2	-0.0505	-0.045	-0.0429
	Bias ρ	0.009	-0.0068	-0.0219
	Minutes	0.3241	0.3376	0.3381
	AUC	0.8568	0.8619	0.8753
	n = 6400			
2000 GIBBS	Bias β_1	0.1523	0.1508	0.1344
	Bias β_2	-0.0473	-0.0498	-0.0196
	Bias ρ	-0.0043	0.0019	-0.0267
	Minutes	5.3453	5.4326	5.5655
	AUC	0.856	0.863	0.8744
2008 LGMM	Bias β_1	0.1041	-0.0003	-0.2775
	Bias β_2	-0.0354	-0.0112	0.0484
	Bias ρ	0.0095	0.042	0.1308
	Minutes	3.4408	3.398	3.5013
	AUC	0.856	0.8628	0.8651
2017 ML	Bias β_1	0.1467	0.1637	0.217
	Bias β_2	-0.0472	-0.0492	-0.0407
	Bias ρ	0.0039	-0.0044	-0.0196
	Minutes	4.1482	4.2033	4.2848
	AUC	0.856	0.863	0.8744

Therefore, as a conclusion, there are no relevant differences in precision or bias of the estimator between ML and GIBBS.Both algorithms seem to be ideal for estimating the spatial probit for any sample size or degree of spatial dependence. Only a significant difference in time which, to be fair, can be adjusted through the number of draws in GIBBS. The analyst must decide between all these methods described taking into account these technical features. Currently, we can say that the estimation of the spatial probit has reached a consolidated and robust level for its use in science. The inconvenience of computational cost has been alleviated by the development of these highly optimized algorithms. Furthermore, it seems that the econometric world in practice, is going towards

Cloud Computing which provides instant access to different levels of RAM or number of CORES. It is, without a doubt, the perfect combination for a boom in the field.

1.7 State of art focused on Human-Behavior

Another of the contributions of this first chapter is to indicate the progress of research using spatial probit models. The algorithms for finding centered and efficient parameters are sufficiently robust as seen in the previous section. Although, there are still two factors that go against research with these algorithms. The first, computing times are still high especially when the databases are large. And second, machine learning algorithms are being developed producing very deep and accurate estimates and, despite being non-interpretable, they are attracting the interest of a multitude of data scientists in recent times.

The following paragraphs show the contributions of the spatial probit particularly in consumer human-behavior research area. Our intention is not to make an exhaustive list but to enumerate many relevant works on this topic. This field of research is supposed to be quite broad, however, the low number of publications on the subject using spatial econometrics is appreciated. Haghani et al. (2021) propose an in-depth analysis on the diffusion of discrete choice econometric models. In fact, exponential growth in matter is revealed throughout the 21st century. However, spatial phenomena are hardly considered in this analysis. It seems that research on quantitative methods to solve spatial probit regression is not currently accompanied by a growing number of findings. Billé and Arbia (2019) highlighted this fact by focusing on the health-themed papers. In this latter paper, the need for the application of spatial probit models is emphasized due to their potential uses in the field of health. One of the main objectives of this chapter is to extend the search carried out in Billé and Arbia (2019) to discrete choice human-behavior studies.

In 1973 Danniel MacFadden stated that "A fundamental concern of economics is understanding human choice behavior" (McFadden et al., 1973). Since then, the interest of researchers in modeling human behavior has been increasing. Following the research of Haghani et al. (2021), there are four themes around choice modeling. Apart from the purely methodological, the papers focus mainly on health, transport, consumption or environmental issues. Health-related papers predominate over other topics. The weight of spatial contributions on this topic are insignificant and it is analyzed in Billé and Arbia (2019). The lack of use of spatial econometric techniques can lead to suboptimal models in which the estimators are inconsistent and therefore cannot be extrapolated to other use cases. Furthermore, in the case of estimation of variables' marginal effects on the endogenous, there will be a latent bias caused by not having used the correct specification of the model.

Transportation also receives special interest in academic literature. The type of vehicle used, the distinction between public / private transport or the route of the traveler focuses analyst's attention, see e.g. Baidoo and Nyarko (2015), Fiorio et al. (2013), Antonini et al. (2006). However, although spatial variables such as neighborhood or area are present in these studies, the interdependence of the observations is not captured in the models. Wang et al. (2015) propose a spatial autoregressive probit model in which the propensity to use the bicycle as a form of transport is determined. In a similar line of research, Goetzke and Andrade (2010) proposed a walkability model in New York City, in which the neighborhood effect is shown. Both models show the technical superiority of the probit autoregressive model and show the need to resort to these techniques for the implementation of social and urban policies. The modeled reality indicates that both the use of bicycles and the habit of walking become more attractive in those areas where there are already more users of bicycles or walkers.

Within the title of "environmental" is the subtopic "land use" which is the one that currently has the most spatial references. In 2002, right at the time when algorithms for solving spatial probit regression were taking off, Holloway et al. (2002) published a study explaining the Bayesian algorithm (Metropolis-Hastings) and putting it into practice to explain the adoption of High Yield Variety (HYV) rice in Bangladesh. Until then, there had only been studies in which the problem was solved through the classic OLS without considering spatial dependence. As a conclusion to the article, the authors present the model with and without neighborhood effects, where the bias in the marginal effect of certain variables is exactly appreciated, leading to different conclusions between both models. After this analysis, several papers came up analyzing different technological proposals for agriculture (Nazari Gooran and Borimnejad, 2015, Ommani and Noorollah Noorivandi (2019), Skevas et al. (2021), Zheng et al. (2021)). In all of them spatial probit models are specified and Bayesian framework is used for estimation. What is clear is that neighborhood attitudes affect individual decision-making in agricultural land use. Other references related to land use analyze, for example, deforestation in Latin American countries. While (Arima, 2016) uses spatial methods to conclude which lands are more likely to be deforested in the future depending on the distance to nearby roads and the probability of deforestation from other adjacent lands. Other very present methodologies to approach the problem are related to non-interpretable "machine-learning" techniques. For example, Valle et al. (2020) establishes a random forest model and another generalized additive models to find a solution to the problem. As far as we know, there is still no comparison between methods with assembled models versus spatial methods in terms of precision and fit of the residuals.

The last group of papers belong to pure consumer behavior. It is reasonable to think that behavior of individuals is marked among other factors by unobservable variables related to space, by word of mouth or by social movements or trends within each neighborhood. Therefore, the idea that each person forms their own preferences and based

on these, makes their decisions is quite unrealistic. Yang and Allenby (2003a) demonstrate these mimetic behaviors in the choice of car brand. The authors even indicate that behavior influenced by proximity is more significant than due to demographic factors. However, socioeconomic and demographic variables are much more common in the consumer econometric literature. Giansoldati et al. (2020) carried out an interesting analysis on the choice of electric vehicle versus fuel. One might think that environmental concern could impact in this decision and complex spatial patterns could be behind this climate thoughts. If spatial patterns exist, the estimated coefficients could be masking in some way the spatial autocorrelation, causing that those decisions made, based on model results, might not be correct. In this sense, a widespread problem is the customer churn detection and possible actions to retain, with a bast academic literature on this matter. De la Llave et al. (2019b) is the first paper in which a spatial econometric model is specified to detect customer churn. In this case, the model gains slightly in accuracy, but what is more relevant is the change in the marginal effects on the endogenous variable after using a SAR model. Ferreira et al. (2019) also raises this issue by focusing its analysis on customer churn associated with the behavior of each individual's network of friends. With this new approach, the calculation of the customer's lifetime changes and therefore the commercial strategy to retain customers changes. Similar strategies to build customer loyalty in retail ecommerce are included in a spatial study in de la Llave Montiel and López (2020). In many cases, the loss of the customer hides a loss of competitiveness against their peers. Technological advances are a key piece in the market. These advances are not random in space but often depend on zonal behaviors or imitative behaviors of SMEs (Autant-Bernard et al., 2007). The ability to handle geo-referenced data is increasing rapidly and given the need to understand customer churn patterns to create tailored actions, spatial techniques for extrapolation of micro-territorial data becomes a fundamental tool to better approach the consumer behavior.

Let's finish the state of the art by looking for the research that goes after the real reasons why humans perform mimetic behaviors to those around them. As we have seen above, there is mathematical evidence that models that take into account proximity between individuals have better properties in the investigation of human behavior. This is simply the evidence of some much deeper biological, psychological or neuronal process. Mirror neurons are cells found in the premotor cortex of the human being. According to Rizzolatti and Craighero (2004), these neurons are activated in the cortex of the brain by observing the performance of an action by another person. Recent research indicates that imitation as social mirroring requires the connection between the core circuitry of imitation and the mirror neuron system (Iacoboni, 2005). The mirror neural system is responsible for transforming visual information into knowledge. This type of learning is very useful for the neural system since it avoids consuming time and energy in carrying out different tests to know the result. Therefore, in addition to pure knowledge, another aspect in which mirror neurons play a key role is in accepting the result of the actions

performed by others through pure visualization (Cattaneo and Rizzolatti, 2009). When these actions made by a third person generate an outcome, it automatically feeds the observer's decision-making system. From a psychological point of view, imitative behavior is a form of social communication. Through the imitation of behaviors there is a greater connection between people and feelings of empathy are shared (Chartrand and Bargh, 1999). Moreover, this communicative exchange is stronger when it is perceived to be being imitated by another (Meltzoff and Decety, 2003). The reality is that there is a very extensive literature on the neurological aspects of imitation. From the unconscious behavior of infants to the most rational of adults. Currently there is an open line of research to understand what mechanisms are activated in the brain in this complex process. These processes can be the biological basis of what is found in spatial econometric models: Human behavior can be better explained by knowing people's surroundings.

Chapter 2

The impact of geographical factors on churn prediction

2.1 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 receiving increasing interest 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 den Poel, 2012; Millo and Carmeci, 2011); in relation to customer churn behaviour, however, no research that takes into account geography and 'space' as explicative factors has yet been undertaken.

Customer churn 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 declined, 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., 2012) 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 prospective customers (Saradhi and Palshikar, 2011). Therefore, predicting policy

cancellation before the end date is a critical point for most companies. If those groups of customers or policies can be detected, wherever the risk of churn is high, specific marketing actions (e.g. customer retention programs) can be developed in order to keep the customers. A small decrease in retention rates should therefore provide the company with benefits; it is clear that customer retention is a critical point 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 K"opcke, 2004), subscription services (Coussement and den Poel, 2008), financial services (Larivière 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 use classification trees (Lemmens and Croux, 2006) and logistic regression (Günther et al., 2014); multiple statistical techniques[1] have also been developed in order to identify customers who are likely to churn based on their characteristics: for example, survival analysis (e.g. Brockett et al., 2008); neural networks (Hung et al., 2006); random forest (Larivière and den 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 when churn modelling.

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 sociodemographic characteristics (education level, age, sex, family size, social status) and specific information about the company's relation with the customer (number of policies, discount program), and even relationships with another customers (social network, family relations between customers). There is one piece of information, which is included in all data banks 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 know the neighbourhood (or zip code) of the customer, then indirectly you can gain information about the customer's economic status, and it allows you to divide customers into exclusive neighbourhoods. Some research that has been undertaken on zip codes in churn prediction (Lochl et al., 2009; Verbeke et al., 2011; Huigevoort and

[1] 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. (2012); Table 1 in Soeini and Rodpysh (2012) ; Table 1 in Keramati 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).

Dijkman, 2015) showed ambiguous evidence. Verbeke et al. (2011) 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 with our 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 probably is codetermined, and some mimetic conduct between them can be noted. 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. 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 of the 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 (Lochl 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, p. 447).

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 a portfolio of private insurance customers from a major Spanish company, we will prove the power of using geographical information to improve the classical probit regression models using Spatial Regression Probit Models (LeSage and Pace, 2009b). We selected this methodology based on the Spatial Autoregressive Probit model because the usual probit model is a popular methodology that has been shown to perform well in churn analyses. Moreover, the parameter estimates are easily interpretable. We would like to highlight that the 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 chapter 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. The last section concludes this work.

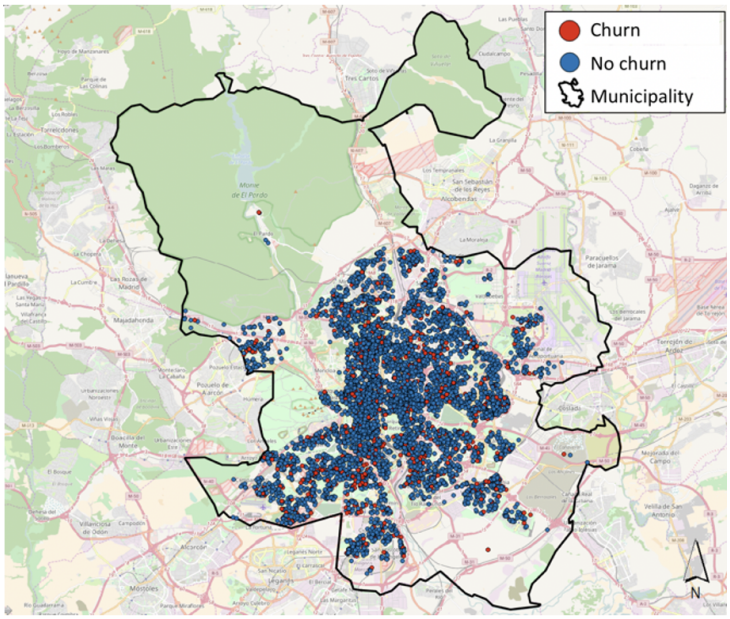


Figure 2.1: Georeferenced customers in the urban area of Madrid

2.2 Data and methodology

2.2.1 Data

The information used in the analysis comes from a large Spanish insurance company which provides a wide range of insurance lines. Data analysed in this paper refer to those customers that have taken out at least one policy through the company's branches. Customers who contracted their policies directly are excluded because their behaviour tends to be different. For this work, we selected only customers located in the municipality of Madrid (Spain). 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. 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 (geocode function in Kahle and Wickham, 2013). Observations non correctly localized or poor geocoding addresses were excluded from the sample. As a result of the aforementioned filters, the final dataset consists of a sample of 7,302 customers. Additionally, the addresses of all insurance agencies of the analysed company (a total of 114 offices) and the more relevant insurance competing agencies (a total of 250 offices) were obtained. The coordinates of all agencies were obtained using similar procedure. Figure 2.1 shows the analysed urban area and the spatial distribution of cases and agencies.

A customer churns if he leaves the company, cancelling all his policies with the company. The overall lapse for the whole sample portfolio is 11.8% which is, according to statistics published by the insurance institute ICEA , similar to figures for other companies in

the Spanish insurance market. These cancellations mainly consist of non-payments or voluntary surrenders. Reasons for cancellation are unknown because there is no any further questionnaire for the relevant customers. 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 socio-demographic characteristics of customers and the contractual terms of their policies. Also, using information about the exact localisation of customers and agencies, we noted other geographical variables that we think could be relevant in this research. The description of all the analysed variables can be found in Table 2.1.

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, we pay particular attention to the duration of the customer-company linkage (Years), the number of active policies that the customer had 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 (68%) have only one policy with the company. 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 competence agency (Dist-Compet). The average distance to an analysed company agency was 964 meters whereas the average distance to a competence agency was 628 meters.

2.2.2 Methodology

Discrete choice models are popular tools to explain the effects of various factors on 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:

Table 2.1: Description of the variables and descriptive statistics

Variable	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 if the customer has no children	0.25 (0.43)	0/1
Age	Age of the customer	49.7 (12.2)	[17-75]
Contractual terms			
Policies	Number of policies with the company	1.52 (0.96)	[1-9]
Years	Number of years as customer of the company	5.23 (2.73)	[1-10]
Premium	Total premium (in thousands €)	0.48 (0.52)	[0.04-6.47]
Geographical variables			
Dist-Own	Distance in meters to the nearest analysed company office	964 (732)	[0-7351]
Dist-Compet	Distance in meters to the nearest office of competition insurance company	628 (444)	[0-7397]

$$y_i := \begin{cases} 1 & \text{if } \; Y^* > 0, \\ 0 & \text{if } \; otherwise \end{cases} \tag{2.1}$$

Note the difference in unobservable profits associated with the 1-0 choice indicators: $\eta_{1i} - \eta_{0i}$, 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:

$$y_i = \begin{cases} 1 & \text{if } \; y_i^* = \eta_{1i} - \eta_{0i} > 0 \\ 0 & \text{if } \; y_i^* = \eta_{1i} - \eta_{0i} \leq 0 \end{cases} \tag{2.2}$$

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:

$$y_i^* = \gamma_1 + \gamma_2 x_{2i}^{ns} + ... + \gamma_k x_{ki}^{ns} + + \epsilon_i^{ns} \tag{2.3}$$

In matrix term,

$$Y^* = X^{ns}\gamma + \epsilon_{ns} \tag{2.4}$$

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, $\epsilon^{ns} = N(0,1)$ $(\forall i)$. 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 2.1 (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 (2001b; 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.

2.2.3 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 2.1: 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:

$$y_i^* = x_i^{ns'}\gamma + x_i^{s'}\delta + \epsilon_i \tag{2.5}$$

In matrix terms,

$$Y^* = X\beta + \epsilon \tag{2.6}$$

Model (3) is nested in the model expressed in (2.5). However, from an estimation and interpretation point of view, they do not present any difference. For the notation of model (2.5), customer i leaves the company with probability P_i, which can be expressed as follows:

$$P_i = P(y_i = 1|x_i) = \Phi(E[Y_i^*]) = \Phi(x_i'\beta) = \int_{-\infty}^{x_i'\beta} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{2.7}$$

where $\Phi(.)$ 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.

Estimation of the model was carried out using maximum likelihood, which in this case deals with the search for those values of the β parameter vector that maximize the log likelihood function, which is expressed as follows:

$$l = lnL(\beta) \ = \sum_{y_i=1} ln\Phi(x_i'\beta) + \sum_{y_i \neq 1} ln(1 - \Phi(x_i'\beta)) \tag{2.8}$$

Next, from the maximum-likelihood estimator, $\hat{\beta}$, mainly two types of results can be noted (Scott Long, 1997; Franses and Paap, 2001, among others). Firstly, the likelihood of the churning of a customer i can be estimated using estimator vector , as follows:

$$\hat{P}_i = P(y_i = 1|x_i) = \Phi(x_i'\beta) \tag{2.9}$$

Secondly, the marginal effect of all explicative variables in the model can be estimated. In the case of a non-factor variable x_h, the marginal effect associated with a customer i refers to the change in the consumer's expected probability of churning due to an infinitesimal change in the variable, as follows:

$$\frac{\partial P_i}{\partial x_{hi}} = \frac{\partial \Phi(x_i'\beta)}{\partial x_{hi}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i'\beta)^2}{2}} \beta_h = \phi(x_i'\beta)\beta_h \tag{2.10}$$

where $\phi(.)$ denotes the standard normal density. As is also commonly the case in literature, we use (2.10) as an approximation for the change in probability of churning produced by a one-unit change in the variable. Furthermore, from (2.10) it can be deduced that changes in the value of the variable for another customer j, x_{hj} do not influence customer i's decision.

A general conclusion related to the marginal effect of the variable x_h itself, $\frac{\partial P}{\partial x_h}$, will relate to the average of all customer marginal effects.

By contrast, the marginal effect of a factor variable, said to be x_f, is defined as the difference between the average probability of churning for the two possible values of the variable (1/0), as follows[2]:

$$\frac{\partial P_i}{x_f} = average(P\{y_i = 1|x_{fi} = 1\}) - average(P\{y_i = 1|x_{fi} = 0\}) \tag{2.11}$$

2.2.4 Type II spatial probit

Next, for a **type II spatial probit** proposal, we will discuss the Spatial Autoregressive (SAR) probit model, proposed by LeSage and Pace (2009b). Following the notation for model (2.5), it reads as follows:

$$Y^* = \rho W Y^* + X\beta + \epsilon;\ \epsilon = N(0, I_N) \tag{2.12}$$

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, the row-standardization 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 (10), 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 measures the strength of dependence, with a value of 0 indicating independence. Clearly the first type of spatial probit model emerges when

The reduced form of expression (2.12) can be written as follows:

[2]Another possibility is the evaluation of the density term in (2.10), $\phi(x_i'\beta)$, at the mean values of all regressors. In this case, the marginal effect associated with variable x_h is interpreted as the change in the probability of churning associated with a change in the average (or typical sample observation) of such a variable. Large N results are similar, although some differences can appear for small sample sizes.

$$Y^* = S(\rho)X\beta + S(\rho)\epsilon \tag{2.13a}$$

$$S(\rho) = (I_N - \rho W)^{-1} = I_N + \rho W + \rho^2 W^2 + ... \tag{2.13b}$$

$$X\beta = \sum_{h=1}^{k} x_h \beta_h \tag{2.13c}$$

while the vector of probabilities of leaving the company can be obtained as follows:

$$P = P(y = 1|x) = \Phi(E[Y^*]) = \Phi(S(\rho)X\beta) \tag{2.14}$$

The SAR probit model has been revealed to be very useful in the fields of economics, political science, sociology, ecology, planning and even neurology. Nevertheless, its complexity has resulted in a reduced number of applications in comparison with those generated on the basis of the SAR continuous dependent variable model. Most of the applications refer to the choice between alternatives: for instance, whether or not to adopt a new farming technology (Case, 1992), increase tax rates in a district (Beron and Vijverberg, 2004b), reopen damaged infrastructure (LeSage et al., 2011), where to locate plants (Klier and McMillen, 2008; Collingham et al., 2000), R&D labs (Autant-Bernard, 2006), harvest trees (Fortin et al., 2013), defoliate (Heagerty and Lele, 1998) or deforest (Brun et al., 2015).

Of the spatial SAR probit estimation proposals in the available literature, the Generalised Method of Moment (GMM) estimators proposed by Pinkse and Slade (1998) and Klier and McMillen (2008) will now be discussed. They are derived from a weighted nonlinear version of the linear probability model associated with the observed variable. According to Calabrese and Elkink (2014), although the GMM estimators present a high computational speed, they suffer from poor accuracy, especially when the value of the spatial dependence parameter is high. Alternative proposals are derived from the underlying latent equation expressed in (2.12). Within this group, we should note McMillen's proposal (1992), which provides estimates by means of the Expectation-Maximization (EM) algorithm proposed by Dempster et al. (1977b). However, the main drawbacks of this estimation procedure are the imprecision of the estimation of the spatial autoregressive parameter, and its impracticability for large sample sizes. Next, the application of the Bayesian Gibbs sample approach to the SAR probit model, which was proposed by LeSage (2000b), solves most of the issues with McMillen's proposal. According to Calabrese and Elkink (2014), the Bayesian Gibbs estimator performs reasonably well in term of accuracy, but its use is unfeasible for large samples (N $>>$ 1000). Finally, other works use Maximum Likelihood (ML) procedures, mainly because of the advantages associated with, for instance, the use of Likelihood Ratio (LR) tests. Next, we proceed by analysing a little further the main ML estimation proposals in the available literature.

Let us now define the error term of the reduced expression in (11) by $\nu = S(\rho)\epsilon$. Taking into account that $\epsilon = N(0, I_N)$, the variance of the error term can be written as follows:

$$\Sigma = E[\nu\nu'] = S(\rho)S(\rho)' \tag{2.15}$$

To obtain consistent and efficient estimates of the β and ρ parameters using the ML procedure, it is necessary to evaluate the following N-dimensional normal probability:

$$l = lnL(\beta, \rho) = \Phi_N(x \in A|\Sigma) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} \int_{A_1} \int_{A_2} ... \int_{A_N} e^{-(1/2)x'\Sigma x} \tag{2.16}$$

where $A = \{A_i\}_{i\in\{1,2,...,N\}} = (a_i, b_i)_{i\in\{1,2,...,N\}}$

being

$$a_i = \begin{cases} -\infty & \text{if } \ y_i = 1 \\ S(\rho)X\beta & \text{if } \ y_i = 0 \end{cases} \tag{2.17}$$

and

$$b_i = \begin{cases} S(\rho)X\beta & \text{if } \ y_i = 1 \\ \infty & \text{if } \ y_i = 0 \end{cases} \tag{2.18}$$

In this context, Beron and Vijverberg (2004b) proposed a Recursive Importance Sampling (RIS) estimator, which allows for the evaluation of N-dimensional probit likelihood. As in the case of the Bayesian Gibbs, the RIS estimators perform reasonably well in term of accuracy, but their use is unfeasible for large samples (Wang et al., 2011; Calabrese and Elkink, 2014). To solve this problem, Pace and LeSage (2017) and Martinetti and Geniaux (2017) proposed to work on sparse variance-covariance matrices or the sparse precision matrix (the inverse variance-covariance) to make this procedure feasible with large bodies of data. However, while Pace and LeSage's proposal (2017) uses the RIS simulator (as Beron and Vijverberg, 2004b), Martinetti and Geniaux (2017)'s procedure is based on a modified version of the Mendell and Elston approximation method (Mendell and Elston, 1974a). The procedure developed by Martinetti and Geniaux (2017) is able to approximate the full log-likelihood function, although they conclude that the use of conditional log-likelihood (CL) is very efficient and reliable since conditional estimators outperform the respective full-likelihood estimators. The CL uses intermediate estima-

tions of the covariate parameters, which are conditional on the value of the spatial parameter, to improve the value of the log-likelihood. The application of CL estimations requires a decision be made on whether to operate on the variance-covariance matrix of the likelihood function (UC) or on the precision matrix (UP), which is usually sparser and leads to faster computations. These alternatives give rise to the two conditional estimation versions, named by authors as CLUC and CLUP. Among both estimators, the CLUC estimator presents a higher level of accuracy at the expense of a lower estimation speed.

2.3 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 2.2 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, Policies, Year, Dist-Compet). 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 number of policies in the company, 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.

Secondly, taking into account the results depicted in Figure 2.2, the first step in the specification process consisted of choosing 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 for the baseline model (the non-spatial probit model), considering that socio-demographic factors and contractual policy variables are the only determinants of customer churning. The first column in Table 2.2 describes the results.

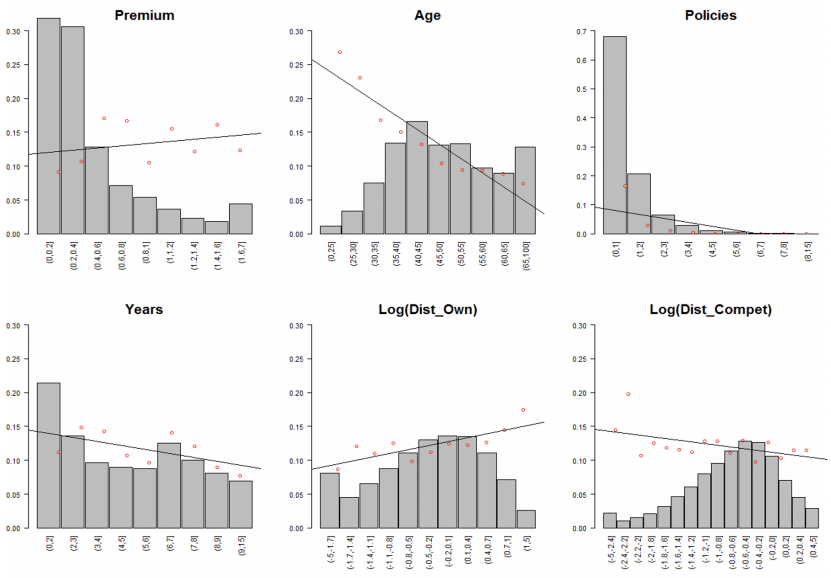


Figure 2.2: Lapse rates and histograms (in segments) for the continuous explicative variables.

Table 2.2: Probit and spatial probit: churn prediction (a), (b)

	Non-spatial probit	Spatial probit	
	Baseline Model	Type I spatial probit	SAR probit model
	Coeff (z-value)	Coeff (z-value)	Coeff (z-value)
Intercept	−1.420*** (-13.2)	−1.448*** (-12.8)	−1.157*** (-10.8)
Socio-Demographic variables			
Gender	−0.253*** (-5.9)	−0.250*** (-5.8)	−0.251*** (-5.8)
Children	-0.038 (-0.7)	-0.040 (-0.8)	-0.037 (-0.7)
h(46-Age)	0.028*** (6.7)	0.028*** (6.7)	0.028*** (6.8)
h(Age-46)	-0.005 (-1.6)	-0.004 (-1.5)	-0.004 (-1.5)
Contractual terms variables			
h(2-Policies)	1.088*** (15.0)	1.089*** (14.9)	1.096*** (15.0)
h(Policies-2)	-0.417** (-3.2)	-0.417** (-3.2)	-0.424** (-3.3)
Years	-0.004 (-0.5)	-0.004 (-0.5)	-0.005 (-0.6)
h(0.549-Premium)	−1.467*** (-10.3)	−1.455*** (-10.2)	−1.470*** (-10.4)
h(Premium-0.549)	-0.029 (-0.5)	-0.016 (-0.3)	-0.027 (-0.5)
Geographical Variables			
Log(Dist-Own)		0.093** (3.1)	0.041** (2.2)
Log(Dist-Compet)		-0.122* (-2.5)	-0.056** (-2.4)
ρ			0.215*** (11.4)
Diagnostic tests of spatial dependence			
I Moran(c) (W10nn)	84.16***	72.85***	
I Moran (W15nn)	61.37***	49.33***	
I Moran (W20nn)	47.50***	33.63***	
Diagnostic tests			
H–L test(d)	0.858	0.902	0.307
AIC	4651.2	4643.3	4634.54
LogLik	2315.6	2309.8	2305.3
LR test		10.6***	9.0***
AUC	0.753	0.755	0.756

(a) h(c-X)={max0,c-X}; h(X-c)={max0,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.

Specification diagnostics for the estimated model are shown at the bottom of Table 2.2. First, the area under the ROC curve (AUC) indicates that the model correctly predicted 75.3% of the choices, and therefore the baseline model exhibited an acceptable level of predictive performance. Furthermore, the null of a correct model specification, tested using the Hosmer-Lemeshow (H-L) statistical test, cannot be rejected at a 5% level of significance. 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 €549 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 become less likely to move to another company and thus the probability of churn begin to decrease. In the same context, as the number of policies a customer holds in the company increases, so 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 for expensive and/or higher policies. We should also note the two slopes; the steeper slope covers the group of customers who hold just one policy, and a softer slope is related to those customers who hold two policies or more. 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 hypothesize 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, 2001b). 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 2.2 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 the effect of geographical distance on the customer and insurance agencies (the analysed company's own and the competition's). The results, which are in the second column of Table 2.2, are clear: the insurance company is more likely to lose a customer if he/she is located near a competence agency or he/she is far from an analysed company agency. The 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 2.2), which is denoted "spatial autoregressive coefficient". Results for the proposed SAR probit model, considering the effects of the customer's $k = 15$ nearest neighbours, are detailed in the last column of Table 2.2. We use the ML methodology propose by Martinetti and Geniaux (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. 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 socio-demographic and contractual variables denotes a high level of robustness in our results.

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).

2.3.1 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 (as shown in Section 2.2.1) while requires more care in the case of the SAR probit model expressed in (10) (Lacombe and LeSage, 2015). 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, $\partial P_i/\partial x_{hi}$, 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, $\partial P_i/\partial x_{hj}$, 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.

The (NxN) matrix of the own- and cross-partial marginal effects, associated with changes in the variable x_h, can be obtained following the expression:

$$\frac{\partial P}{\partial x_h'} = \frac{\partial P}{\partial x_{h1}} \frac{\partial P}{\partial x_{h2}} ... \frac{\partial P}{\partial x_{hN}} = D(\phi(\eta))S(\rho)I_N\beta_h \tag{2.19}$$

where $D(\phi(\eta))$ is an (NxN) diagonal matrix (with zeros outside the main diagonal). The whole main diagonal vector will be denoted as follows: $d(\phi(\eta)) = [d(\phi(\eta_1)), d(\phi(\eta_2)), ..., d(\phi(\eta_N))]'$, where $\phi(\eta_i)$ element represents the probability density function evaluated at the predictions for consumer i.

Next, we can expand the expression into component matrices gathered in (11), obtaining the following expression:

$$\frac{\partial P}{x_h'} = [D(\phi(\eta)) + \rho D(\phi(\eta))W + \rho^2 D(\phi(\eta))W^2 + ...]\beta_h \tag{2.20}$$

from which, due to the definition of **W** as a row-standardised matrix, the $Nx1$ vector of (cumulative) total effects can be written as follows:

$$\frac{\partial P}{x_h'} = [D(\phi(\eta))\iota_N + \rho D(\phi(\eta))W\iota_N + \rho^2 D(\phi(\eta))W^2\iota_N + ...]\beta_h = \tag{2.21}$$

$$= D(\phi(\eta))\iota_N(1-\rho)^{-1}\beta_h = d(\phi(\eta))(1-\rho)^{-1}\beta_h \tag{2.22}$$

A scalar summary measure of *Average Total Effect* (ATE) is calculated as the average of the vector of (cumulative) total effect:

$$ATE = \frac{d(\phi(\eta))(1-\rho)^{-1}\beta_h}{N} \tag{2.23}$$

To summarize the *Average Direct Effect* (ADE), following LeSage et al. (2011), the following expression is used:

$$ADE = \frac{1}{N}tr\left(\frac{\partial P}{x_h'}\right) = \left[tr[D(\phi(\eta))] + \rho tr[D(\phi(\eta))W] + \rho^2 tr[D(\phi(\eta))W^2] + ...\right]\frac{\beta_h}{N} \tag{2.24}$$

where the efficient computing of the term $tr[D(\phi(\eta))W^p]$ for $p = 1, 2, \ldots$ can be carried out following several approaches proposed in LeSage and Pace (2009b).

Finally, the (cumulative) *Average Indirect Effect* (AIE) is derived by calculating the difference between the previously mentioned effects, that is:

$$AIE = Average\ Total\ Effect\ -\ Average\ Direct\ Effect \tag{2.25}$$

Following this methodology, Table 2.3 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 an 18% higher chance of losing a customer if he/she has just one policy. Cross-selling is vital for companies, as it means good benefits and also because it makes customers more loyal. Consistent with this, once the customer has more than one type of insurance, every new policy signed reduces the direct probability of his/her churning by 7.3%.

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 0.63% per year). This is clearly information which should be used to modulate and optimize the premium renewal every year depending on the personal features of the customer.

Premium is also a variable in the process of optimization. In our sample, we noted that up to €549, every €100 euros paid by the customer increases the likelihood of lapses by 2.5%.

Table 2.3: Direct, indirect and total effect of the spatial probit model

	Direct effect	Indirect effect	Total effect
Socio-Demographic variables			
Gender	-0.0434	-0.0117	-0.0551
Children	-0.0064	-0.0017	-0.0081
	h(46-Age)		
h(Age-46)	-0.0008	-0.0002	-0.0010
Contractual terms variables			
h(2-Policies)	0.1897	0.0514	0.2411
h(Policies-2)	-0.0734	-0.0199	-0.0933
	Years		
h(0.549-Premium)	-0.2544	-0.0689	-0.3233
h(Premium-0.549)	-0.0047	-0.0013	-0.0059
Geographical Variables			
Log(Dist-Own)	0.0071	0.0019	0.0090
Log(Dist-Compet)	-0.0097	-0.0026	-0.0124

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 2.3, it is worth using geographic variables in the analysis, as it gives a sense of how dominant branches are with regard to the competition. Thus, 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 (kilometer) of distance among customers and tied-agent increases the probability of churning by about 0.9%. This likelihood might be increased if competence has closer branches. An additional log (kilometer) of distance between customers and the competence reduces the churn probability to 1.2%.

2.4 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 may 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 neighbour. 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 (2009b)).

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 (kilometer) of distance between customers and a company tied-agent increases the probability of churning by about 0.9%. An additional log (kilometer) of distance between customers and the competence reduces the churn probability by 1.2%. 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 with 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 €549, every €100 paid by the customer increases the likelihood of lapses by 2.5%. As for the number of policies, the results indicate that the chances of losing a client increase by 18% if he has just one policy.

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 georeferencing work on business premises needs to be undertaken. Lastly, in this paper we have demonstrated the importance of distance (in meters) 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.

Chapter 3

Spatial models in online retailer churn

3.1 Introduction

Online retailing is revolutionizing the retail landscape (Wood, 2011). With the popularity and high accessibility of the Internet, consumers have been actively using online channels for their shopping. The consumers have multiple e-shopping alternatives (fashion, travels, hotels, flights, technology, food, etc) and large companies like Amazon, AliExpress or eBay are a real alternative to shopping. Like the rest of dot-coms, the new format of selling groceries online has grown at a phenomenal pace. Online supermarkets have been gaining popularity among consumers mainly in great urban areas, when the cost of transport, in time, to access to hypermarket or large supermarket is high. The speed and comfort to carry out the buying, without having to go to the physical store, attracts to consumers that demand more free time (Martínez and Vázquez, 2008).

With the objective of increasing the customer portfolio, dot-com e-commerce companies spend large amounts of resources trying to gain new customers each year, but many of them consistently fail to retain such customers. An important topic of research of traditional offline supermarket is the churn rate (Burez et al., 2009). In case of classical supermarkets more than 25% of customers churn (Burez et al., 2009) and close to 90% of grocery shoppers use two or more supermarket for their grocery shopping (Miguéis et al., 2012). In case of dot-com companies the churn rates are higher than the traditional supermarket due to the specific characteristic of this type of commerce that makes having loyal customers an extremely complicated task. It is difficult to retain customers due to increased competition and minimal customer switching costs in the online environment (Srinivasan et al., 2002). Low transportation costs online allow shoppers to visit multiple e-commerce sites for a purchase decision (Park, 2017). In the grocery retail environment engaged customers create higher benefits than new ones (Reichheld and Sasser, 1990) but

the cost of gaining a new customer overpasses the cost of retaining the same customer. Some research has shown that this cost varies between six (Verbeke et al., 2011) and twelve times that of retaining the existing customer (Torkzadeh et al., 2006). So it seems only logical to avoid churning as much as possible in order to create a sustainable business model and also to elude adverse effects such as negative company's reputation or negative feedback which may influence potential customers (Saradhi and Palshikar, 2011). Therefore, developing algorithms to identify the factors related with churn and with power to predicting churning before it happens is a critical point for grocery e-companies. If such group of customers can be detected, wherever the risk of churning is high, specific marketing actions can be developed in order to retain such customers. In summary, building models to explain and predict the churn, is crucial for companies to conduct effective retention campaigns.

Dot-com companies collect large amount of information from their customers, which can be analysed to find certain valuable patterns. For each customer, thousands, or even millions of data objects are stored, enabling the analysis of the complete purchasing history. They store data from the very first moment of the customer using the service. The specific date when the customer signed up and all the events of that specific moment could reveal something interesting about coming customer's behaviour. Date might reveal the economic balance situation of consumer's bank accounts, so it plays a critical role in behavioural purchasing models. In addition, channel of entrance (Facebook, e-mail, etc) is decisive to discriminate between loyal customers to the ones who become inactive in a short period of time Richards et al. (2014). Moreover, first user's orders impact meaningfully on the prediction of future behaviour of a customer, as probably is an indicator of the willingness to buy things on the internet. Both total orders made and variety of basket seem to us very powerful indicator of how linked the customer will be in the future. These variables are commonly used, for instance in Miguéis et al. (2012). But there is one piece of information, which is included in all data warehouses that has high relevance in models of churn prediction: the address of the customer. This information is critical for several reasons.

First, knowing the address it is possible have information about the neighbourhood of the customer, then indirectly the economic position is revealed. This information is a key to predict the probability of churn. Some research that has been undertaken on zip codes in churn prediction (Lochl et al., 2009; Chu et al., 2007; Verbeke et al., 2012; Huigevoort and Dijkman, 2015; Trivedi, 2011) showing ambiguous conclusions. Trivedi (2011) write, "Retailers would benefit from understanding the spatial, demographic and attitudinal effects that play into consumption behaviour, and such effects can be better understood when studying choice at the category and region level". In the opposite direction, Verbeke et al. (2012) wrote, *"the age of a customer turns out to have good predictive power, but zip code or similar information on the other hand not at all, as might be expected"*. But

most of research about spatial customer behaviour that take account the geography is analysed generally grouped into geographical units of variable size (municipalities, zip codes, census tracks). Data aggregation involves some degree of arbitrariness, which may lead to biased results (Modifiable Areal Unit Problem, MAUP). In case of dot-com companies, the information is available at the greatest level of granularity (georeferenced spatial point patterns), because it is necessary to take the purchase home. Then, the question that naturally arises is the reason why data is analysed at group level when same information is available in a disaggregated form, and precisely georeferenced. In this research we will consider the exact localization of customer.

Secondly, and directly related to this research, one important aspect of grocery shopping is distance to the store. If the e-grocery company knows the exact location of a customer, it is easy to identify geographical factors (strategic geographical points) that could be related to churning. Proximity to certain supermarket is probably an essential factor that influences in churn performance (Elms et al., 2016). There is a huge literature about the relationship between grocery shopping and distance (Hsu et al., 2010; Nilsson et al., 2015). We will consider this information as relevant in our research and show how there is a non-linear relationship between churn probability and distance to favourite physic supermarket.

Lastly, by knowing the exact location of a customer (latitude and longitude) 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, rumours, commentaries and mostly activities of churn from 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. The emergence of relational networks is inseparable from social communities. In the analysis of these networks, it must be taken into consideration that the behaviour and beliefs of an individual are generally influenced by the behaviour and beliefs of the rest of individuals, especially those with whom they interact directly. These direct interaction among individuals have been referred by network-related literature in different ways: social contagion (Leenders, 2002); social effect (Manski, 1993); pairwise or neighbourhood effect (Moffitt, 2001); consumer preferences (Yang and Allenby, 2003b); homophily[1] (Zhang et al., 2012). With the evolution of social networks considering social ties in churn prediction has proven to be a promising approach (Droftina et al., 2015) and a few papers consider the network structure to improve the churn models. For example, the number of neighbouring churners and the number of calls to neighbouring churners, or the effect of word of mouth on churn in mobile phone market (Dierkes et al., 2011). Likewise, Yang and Allenby (2003b) stated

[1]Unfortunately, the methodology used in this research does not allow identifying the cause of the interaction between customers and more research is necessary about this topic.

that the geographical interdependence between individuals is important in explaining car consumption. Therefore, taking advantage of the spatial correlation among customers can improve the performance of models. However, in relation to customer churn behaviour, no research that takes into account geography and 'space' as explicative factors has yet been undertaken and only the research of (De la Llave et al., 2019a) include geographical factors.

Although there has been much literature around prediction of customer churning in the last years, none of it takes into account spatial models to do so. So, this paper is based on the belief that spatial econometrics models can improve the results of classical methodologies that do not consider geographical factors.

Taking into account the state of research, the main objective of this paper is to demonstrate the impact of geographical factors on churn models and show the improvement achieved on the prediction of churn. Using the experience of a dot-com e-commerce startup business in Madrid (Spain), it is proved in this paper that the power of using geographical data to improve the classical probit regression model using Spatial-Probit Models (LeSage and Pace, 2009b). This paper proposes a new take on the classic churning prediction analysis, which may lead to an improvement in the business model of companies and resource allocation thereof.

We thus contribute to the literature in several ways. In first place, as far as we know, this paper is the first research that uses the exact localization of customer, as a statistical points process, in an urban environment to improve the churn model. In second place, we focus on the importance of mimetic performance of close customers in the retention process within a directed social network. The paper is structured as follows: the second section describes the data and methodology; the third section presents the most important results and the last section conclude and suggest some potential e-companies' strategies.

3.2 Data and methodology

3.2.1 Data

Our database comes from a dot-com company which provides the user to do the shopping online in his favorite supermarket (chosen by the user) and receive it in a very few hours. This dot-com company is an emerging startup business which operates in some regions of Spain. Despite its small size, it is positioned among the most demanded internet companies among the ones that provide a similar service. In this work, we selected users in Madrid urban area, as this is the city where the firm is mostly consolidated. The georeferencing used to locate these customers is the address given by the customer to receive the shopping, so the exact coordinates (latitude and longitude) were integrated into the

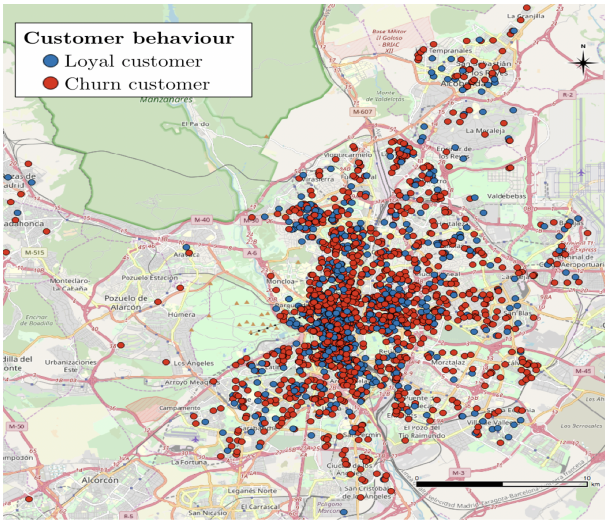


Figure 3.1: Urban area with spatial distribution of individuals

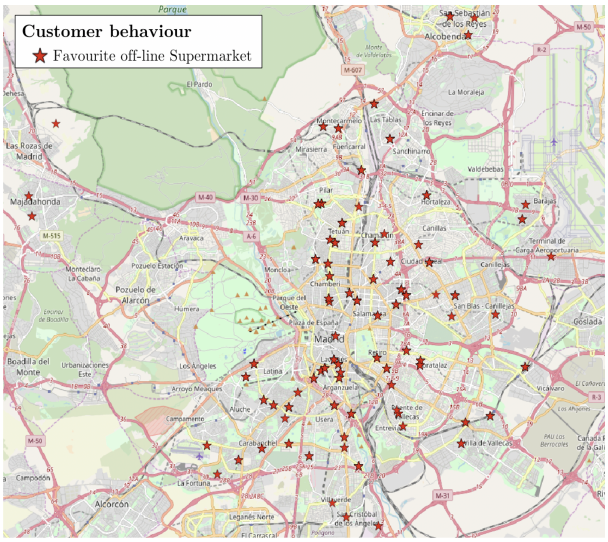


Figure 3.2: Urban area with spatial distribution of favorite supermarkets

database. Figure 3.1 and 3.2 shows the analysis urban area and the spatial distribution of individuals and supermarkets.

Additionally, addresses of supermarkets chosen by users (a total of 84 supermarkets) and a list of all the supermarkets of the city (a total of 586) were included in the geospatial information. Figure 2 shows the localization of supermarkets. All the customers taken into account have ordered a shopping basket at least once during their first week after registration. The customers who get registered into the company and do not start their shopping in the first week are excluded from the database analyzed, although they are not really material as they represent less than 1% of the database received from the company. The period studied goes from January 2016 to February 2017. As a result of the selection, our analysis will be built on 1731 observations to train the model and 306 (15% of the data) observations randomly chosen to test it.

The rule to determine the churning event is applicable to those customers that after their first week experience after registration with the company, remains inactive for at least four months. After this period of time no customer has made an order again, consequently these customers can be considered as lost. Similar approaches have been taken in the literature, for instance, in a similar study Buckinx and Van den Poel (2005), state that "Customers are considered to break their relationship when they interrupt their loyal and stable purchasing pattern that they exhibit during a period of five months", and Lai and Zeng (2014) in a study of churning in libraries found that the churn hazard was in the first three months after registration.

The description of the variables analyzed is depicted in Table 3.1. The overall churning for the whole portfolio is 69.2% (1410 customers out of total 2037). Customer's reasons for inactivation are unknown because there is no any further questionnaire asking for possible triggers.

Table 3.1: Description of the variables and statistics

	Definition	Mean (std)	Range
Dependent variable			
Churn	=1 if customer became inactive.	0.69 (0.46)	0/1
Independent variables			
Channel			
Facebook	= 1 if customer's channel to the App was Facebook.	0.17 (0.38)	0/1
email	= 1 if customer's channel was via Email.	0.17 (0.38)	0/1
Social	= 1 if customer's channel was other Social Network.	0.03 (0.18)	0/1
Socioeconomic			
Male	= 1 if customer's gender is male	0.03 (0.18)	0/1
Login Moment			
Day	Day of the month in first connection.	15.8 (8.79)	1-31
Month	Month of the year in first connection.	15.8 (8.79)	1-12
App	= 1 if customer's logged with mobile App.	0.09 (0.28)	0/1
Temperature	Maximum temperature in area the day of login.	19.4 (8.45)	4-39
First Week Experience			
Orders	Orders during first week using the App.	1.25 (0.57)	1-5
Av-Invoice	Average invoice during first week.	70.8 (47.6)	5-400
Basket	Number of type of products bought in first week.	9.57 (2.92)	1-16
Geographical			
Log-Income	Average income per census track.	6.84 (0.36)	9-11
Orders-Neighbors.	Five closest neighbours' orders average.	1.24 (0.26)	1-3
Log-Distance	Log Distance to the favourite customer's supermarket.	0.07 (0.86)	(-3)-3

The customer's information captured in this type of businesses is very limited, as customers are reluctant to give much information, so for commercial purposes just essential customer's information is required by the company. Our database consists of the gender of the client, information regarding the customer's first connection (date, temperature, device used), the channel whereby they found the dot-com company and the data about

their orders during their first week after the registration in the company. Additionally, the data base is enriched with geographical data regarding customer's home location, distance between nearest supermarkets and also the census track of log-income statistic published by municipal institutions in order to distinguish between different levels of wealth in Madrid.

3.2.2 Methodology: Spatial autocorrelation test for qualitative data

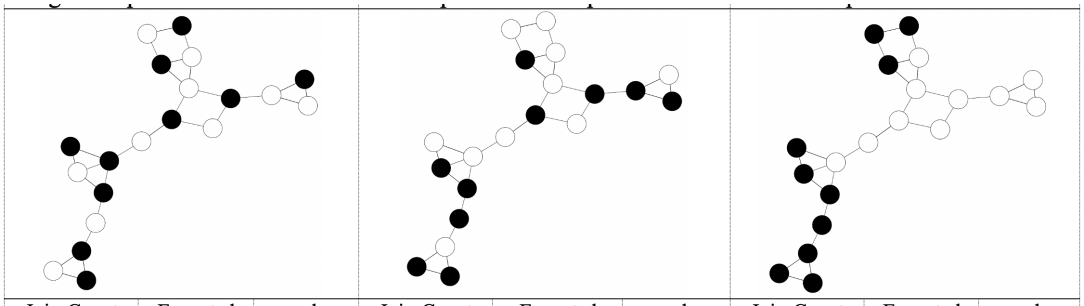
The Join-Count statistic (Cliff and Ord, 1981) will be used to test the null of random co-localized pattern of churn/loyal customers. The Join-Count statistic counts the number of each of the possible "joins" between neighbours. Possible joints are CC (churn-churn), LL (loyal-loyal), and CL (churn-loyal). The statistics J_{CC}, J_{LL} and J_{CL} count the observed number of joins and compare with the expected number under the null (J'_{CC}, J'_{LL} and J'_{CL}). In order to join customers, we selected a binary weight matrix W to establish a connectivity criterion. In particular, the elements of W, w_{ij} (i,j=1,...,N) have a value of 1 if customers i and j are neighbours, and 0 if otherwise. We consider that two customers "i" and "j" are joined if the j-customer belongs to the set of k-nearest i-customer. From this connectivity criterion the Join-Count statistics (J_{CC}, J_{LL} and J_{CL}) are defined as follows in equations:

$$J_{CC} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} CC_{ij} \tag{3.1a}$$

$$J_{CL} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} CL_{ij} \tag{3.1b}$$

$$J_{LL} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} LL_{ij} \tag{3.1c}$$

Spatial co-localised patterns which result from the application of Join-Count tests can be positive or negative. Maps with different spatial configurations of a qualitative variable could be as shown in Figure 3.3 using simulate data. A positive co-localised pattern indicates a spatial structure in which there is a high probability of finding customers that belong to the category C or L surrounded by customers which fall in the same category (Figure 3.3, right), while a negative result reveals the spatial interconnection of customers which fall in different categories (Figure 3.3, left). When the spatial distribution is random, no spatial co-localized pattern can be attested (Figure 3.3, centre). The values of Join-Count statistics including the expected values and z-values are included in the bottom of Figure 3.3. The R package spdep was used to get the statistics.



Join-Count	Expected	z-value	Join-Count	Expected	z-value	Join-Count	Expected	z-value
$J_{LL} = 2.0$	$J'_{CC} = 4.7$	-2.33	$J_{LL} = 4.0$	$J'_{CC} = 4.7$	-0.63	$J_{LL} = 8$	$J'_{CC} = 4.7$	2.78
$J_{CC} = 2.5$	$J'_{CC} = 4.7$	-1.91	$J_{CC} = 4.0$	$J'_{CC} = 4.7$	-0.63	$J_{CC} = 7.5$	$J'_{CC} = 4.7$	2.35
$J_{CL} = 15.5$	$J'_{CL} = 10.5$	2.47	$J_{CL} = 12.0$	$J'_{CL} = 10.5$	0.73	$J_{CL} = 4.5$	$J'_{CL} = 10.5$	-2.99
$J_{CL} = 15.5$	$J'_{CL} = 10.5$	2.47	$J_{CL} = 12.0$	$J'_{CL} = 10.5$	0.73	$J_{CL} = 4.5$	$J'_{CL} = 10.5$	-2.99

Figure 3.3: Urban area with spatial distribution of favorite supermarkets

3.2.3 Methodology: Spatial probit model

Discrete choice models are probably the most frequent methodology used in churn model. In this subsection, we present the extension named spatial-probit models with interdependence in the latent-variable (LeSage and Pace, 2009b) to take into account the spatial effects between close customers.

Let Y be a binary nx1 vector that reflects information on whether or not customers have churned during a certain period; that is:

$$y_i := \begin{cases} 1 & \text{if } \textit{customer i leave company (customer i churns)}, \\ 0 & \text{if } \textit{customer i does not leave company (customer i does not churn)} \end{cases}] \quad (3.2)$$

where Y is the observed value of the limited-dependent variable, $Y*$ is the unobserved latent dependent variable. As a Baseline model, we consider a classical non-spatial probit model, which assumes a lineal relationship between the unobserved latent variable and a set of non-spatial explicative variables:

$$Y^* = X\beta + \epsilon \; ; \; \epsilon \equiv N(0, I_n) \quad (3.3)$$

where X denotes the corresponding (nxm) matrix of covariates and β a $(mx1)$ vector of coefficients.

The disturbance term is used to denote that two customers with the same characteristics can make different choices. If close customers have similar choices, it is possible to find spatial autocorrelation in the residual and therefore in the estimation of (6) we obtained estimated parameters inconsistent and inefficient (McMillen, 1992). The null of

no spatial dependence in residuals of the non-spatial probit model in (5) can be tested by generalized Moran's I test (Amaral et al., 2013). If the null of no spatial dependence were rejected, two specifications would be adequate. In first place improving the model specification by the inclusion of omitted variables related to the geography of the sample (for example, the distance to supermarket or income of the neighborhood of customers as a proxy of the income level. See Table 3.1) and check again the null of independence in residuals. In case of the generalized Moran's I test rejecting the null a spatial-probit model with interdependence in the latent-variable will be adequate. The spatial-probit model is represented as follows,

$$Y^* = \rho W Y^* + X\beta + \epsilon \; ; \; \epsilon \equiv N(0, I_n) \tag{3.4}$$

where the spatial lag of the latent dependent variable $WY*$ involves the nxn spatial weight matrix W[2]. The row-standardisation of the k-nearest neighbour W matrix was adopted in this research. As is well-known, using this approach, the W matrix contains elements of either 1/k or 0. If customer j represents one of the k-nearest neighbours to customer i, the (i,j)th element of W contains the value 1/k. 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 k neighbouring consumers' utility, and it creates a mechanism for modelling interdependence in consumer churn choices. In model 3.4, 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. If $\rho = 0$ the spatial probit model collapses to the standard binary probit model, otherwise, if $\rho \neq 0$, the $(nx1)$ vector $WY*$ consisting of an average of the k churn neighbouring customers and include a mechanism for modelling interdependence in choices. For estimation of spatial-probit model we use the procedure based on Conditional Maximum Likelihood developed by Martinetti and Geniaux (2017). The ProbitSpatial R-package was used to estimate the model.

[2]Alternative models could be selected to take account spatial correlation. The spatial error model is the usual alternative to the model proposes in this paper (the spatial lag model). We follow the paper of Trivedi (2011) who consider that the spatial lag specification is more appropriate, "where the impact on the neighborhood is substantively affected by the adjoining neighborhoods". This is also supported by Baller et al. (2001), who argue that the former specification (spatial lag) suggests a possible diffusion process while the latter (spatial error) suggests omitted variables. Moreover, the Join-Count test has low power for spatial error process.

3.3 Results and discussion

3.3.1 Descriptive statistics

Table 3.1 shows the main descriptive statistics of the dataset. The first remarkable variable kept in the database is entrance channel, as it could determine future customer's behavior. Most of customer's (70%) come into the company either from search engine websites like google without any campaign to promote the App behind. However, directly related to advertising campaigns, Facebook appears to be the channel where more new users are captured (17%), followed by Email referrals (9%). The rest of users started their experience through other specific campaigns in social networks such as Twitter, Instagram, Linkedin and others (4%). The unique socioeconomic variable refers to the gender of the customer which is generated taken the user's name. Female customers represent more than 60% of the portfolio.

The information regarding the first connection to the online grocery is also recorded. A minority of customers used the mobile application to login (10%), the rest connected directly through the company website. With reference to the date of registration it is uniformly distributed both in the day of the week and in the month of the year. Both variables have been treated as continuous in order to detect any time pattern related to for instance, the balance of the user's bank account during the month. Additionally, maximum temperature in the area during the day of registration is included to the data, to check for the potential relationship between weather and user's inactivation.

As our aim is to infer customer's behavior after a week (after registration) of experience with the dot-com services, all the information of user activity in this period of time is used to estimate future inactivity. Most of customers just order once during their first week (80%) spending an average of 70 euros and order nine different types of products (out of 16). Finally, in order to enrich the analysis, locations of the favorite supermarkets declare by the users are included and also other different shopping centers around them. So, 1.7 is the average distance in kilometers calculated between the customer and the favorite supermarket. Also, the effect of number of orders made by closest neighbors of each customer is explored. And the last geographical variable used to predict future inactivity is the income of residential areas analyzed. This variable is published on a yearly basis by public institutions in Madrid and in logarithm goes from one to ten, where ten is the maximum income score. The average scoring for the areas analysed is 6.84.

3.3.2 Spatial co-localized pattern in churn

Table 3.2 shows the values of the three Join-Count tests. Several connectivity criteria based on the k-nearest neighbour (k = 3, 5, 10, 15, 30, 40) will be used in order to identify

the presence of spatial autocorrelation for different number of k-nearest neighbourhood. The results indicate the existence of a positive spatial co-localised pattern in the variable churn. The spatial autocorrelation is higher for k-nearest neighbour lower than k=10 showing that the mimetic performance of customers has a spatial pattern of proximity.

In case of k=3 and k=5, the number of pairs of connected churn customers (J_{CC}) is significantly higher than the expected values (J'_{CC}). These results reveal that churn customers are highly likely to be surrounded by other churn customers. Moreover, the spatial co-localized pattern of customers that belong to different category yields significant results with negative z-value. The observed number of joints CL (J_{CL}) is lower than the expected value (J'_{CL}), showing a spatial structure churn-loyal in our sample. Finally, in the case of loyal customers, the number of loyal customers located in the vicinity of other loyal customers founded is higher than expected.

Table 3.2: Join-Count tests of spatial autocorrelation for churn

	Join-Count test J_{CC}			Join-Count test J_{CL}			Join-Count test J_{LL}		
k	J_{CC}	$J\text{t}_{CC}$	z-value†	J_{CL}	$J\text{t}_{CL}$	z-value†	J_{LL}	$J\text{t}_{LL}$	z-value†
3	1497.5	1463.67	2.22**	1233	1302.66	−3.09***	325	289.17	3.02***
5	2500.5	2439.45	2.98***	2062	2171.10	−3.69***	530	481.95	3.07***
10	4922	4878.90	1.40	4254.5	4342.19	−2.06**	1008.5	963.90	1.96**
15	7387	7318.40	1.70*	6407.5	6513.30	−2.01**	1483	1445.90	1.30
30	14744.5	14636.70	1.58	12840.5	13026.60	−2.41**	2970	2891.70	1.80*
40	19536.5	19515.60	0.24	17281.5	17368.80	-0.96	3922	3855.00	1.26

†The Join-Count statistics are assumed to be asymptotically normally distributed under the null hypothesis of no spatial autocorrelation. Similar results were founded using the bootstrap alternative.

In case of k=10, 20 and 30 the levels of spatial autocorrelation decrease severely. Only a 'repulsion' effect is founded with significant and negative z-value. Finally, for k = 40, the three tests do not reject the null and the spatial process could be considering as random.

Those results show a clear pattern in the spatial distribution of customers and probably some spatial structure must be included in the probit model to make a correct specification of the model by taking into consideration the spatial spillovers.

3.3.3 The classical probit model

In first place, we performed an analysis of the explanatory variables known by the company just after one week logged in the company which would be our Baseline Model. The results coming from the lineal probit model are depicted in Table 3.3. The first part of Table 3.3 describes the results with the relevant split of the explicative variables for our modelling purpose, whilst specification diagnostics for the estimated model are shown at the bottom.

Table 3.3: Probit and spatial-probit in churn prediction

	Model 0: Baseline Model	**Model 1: Linear Probit**	**Model 2: Non-linear Probit**	**Non-linear Spatial-Probit**
	Coeff. (z-value)	Coeff. (z-value)	Coeff. (z-value)	Coeff. (z-value)
Intercept	1.802*** (11.3)	3.477*** (3.82)	3.204*** (3.37)	2.877*** (3.24)
		Channel		
Facebook	-0.161* (-1.95)	-0.145* (-1.77)	-0.132* (-1.71)	-0.114 (-1.40)
Email	-0.800*** (-7.72)	-0.771*** (-7.45)	-0.777*** (-7.50)	-0.779*** (-7.53)
Social	-0.051 (-0.30)	(b)	(b)	(b)
		Socioeconomic		
Male	-0.040 (-0.60)	(b)	(b)	(b)
		Login Moment		
Day	-0.007** (-1.99)	-0.007** (-1.98)	-0.007* (-1.89)	-0.007* (-1.93)
Month	-0.005 (-0.62)	(b)	(b)	(b)
App	-0.220** (-2.08)	-0.220** (-2.09)	-0.218** (-2.07)	-0.219** (-2.09)
Temperature	0.006* (1.65)	0.007* (1.77)	(a)	(a)
		First Week Experience		
Orders	-0.651*** (-10.65)	-0.643*** (-10.92)	-0.649*** (-11.2)	-0.117*** (-11.1)
Av-Invoice	0.000 (0.53)	(b)	(b)	(b)
Basket	-0.033** (-2.51)	-0.032*** (-2.88)	(a)	(a)
		Geographical		
Log-Income	–	-0.136* (-1.70)	-0.138* (1.82)	-0.128* (-1.72)
Orders-Neighbors	–	-0.226* (-1.96)	-0.231** (-1.99)	-0.117 (-1.07)
Log-Distance	–	-0.106*** (-2.93)	(a)	(a)
		Transformed into non-linear (c)		
h(Temperature,25)	–	–	0.022** (2.40)	0.023** (2.47)
t(Basket,10)	–	–	-0.051*** (-3.21)	-0.050*** (-3.16)
t(Log-Distance,0)	–	–	-0.063 (-0.87)	(b)
h(Log-Distance,0)	–	–	-0.138*** (-2.41)	-0.133*** (-2.67)
		Spatial autoregressive coefficient		
ρ	–	–	–	0.129*** (4.39)
		Diagnostic test of Spatial dependence (d)		
I Moran (W05nn)	2.56**	2.53**	2.57**	–
I Moran (W10nn)	1.66	1.70	1.66	–
I Moran (W15nn)	1.00	0.95	0.93	–
		Diagnostic tests		
AIC	2276.05	2257.13	2253.11	2245.52
BIC	2343.48	2319.20	2320.81	2301.72
LogLik	-1126.02	-1117.56	-1114.55	-1112.76
LR test	–	16.9***	6.0**	3.6**
AUC(Train Sample)	0.7131	0.7249	0.7250	0.7253
AUC(Test Sample)	0.6651	0.6789	0.6855	0.7034

* significance at 10%; ** significance at 5%; *** significance at 1% ; (a) Variables transformed into non-linear; (b) Variables removed from model with z-value under abs(1.5); (c) h(X,g)=max{X,g}; t(X,g)=min{X,g], where X is the variable under analysis and 'g' is the breakpoint detected using the GAM methodology; (d) Generalised I Moran test @amaral2013.

The performance of Baseline Model is satisfactory. The results indicate that the selected variables are important when examining customer churn behavior. One of the most relevant variables to explain churning is the customer's entrance channel. Email referrals and recommendations appear to be essential to evaluate the activity of the customer. When the customer enters through this channel, the probability of inactivation decreases sharply, probably due to the selection process that the referral makes when decides to recommend the application. This finding is in line with Richards et al. (2014) research, who evidenced that product recommendation, influences their peers to revise their consumption preference choices. Similar results are found with relation to Facebook entrance whose users become more loyal than customers who came into the application with no known online campaign at all. Other social networks do not appear to be significant to model customer behavior. Richards et al. (2014) also paid attention to these social networks considering them a high valuable marketing tool to influence customer choices. On the other hand, although academic literature indicates that female exhibits higher levels of loyalty Melnyk et al. (2009), gender does not seem relevant in our model. Even though the majority of the population is female, there is no significant difference in probability of churning between them and male.

The moment when a customer signs up in the application might disclose much information about their future actions. The fact that the customer registers through mobile application instead of online website reveals proclivity to use the application service for longer. The results relating to date imply that in warmer days there is more chance of attracting customers less loyal or just for one use. Similar conclusions occur when this happens during the first days of the month. However, the month of login does not seem to be significant to predict further behavior, although this variable is obviously correlated with temperature. Lastly, the user activity during his first week of application's use is vital to understand how involved they would be afterwards. Although the average invoice does not give any insight whatsoever, the number of orders made and the categories of products bought during the first week go in the same direction. There is a strong positive relationship between having both a varied shopping cart and a great deal of orders and remain active after a week registered in the data base of the company. This statement connects with the propensity that people have to purchase things online. By having these variables inside the model, we are implicitly distinguishing people who have experienced online shopping before and those still with different degrees of distrust to it. Finally, the area under the ROC curve (AUC) indicates that the model correctly predicted 70% of the choices, and therefore exhibited an acceptable level of predictive performance.

As previously stated, we hypothesized that geography could also play an important role in consumer churn decisions. 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 generalized I Moran

test. Following the evidences obtain with the Join-Count tests we select the criterion of k-nearest neighborhood connecting each observation with the k nearest (k=5, 10, 15) because the codeterminate behavior has a local effect. The value of this test, displayed at the bottom of Table 3.3 indicates that spatial autocorrelation in the baseline model is noticeable. Therefore, this model should benefit from the use of spatial information.

Consequently, we propose the estimation of a second model, Model 1, taking into consideration variables from Baseline model above 10% significance, in order not to introduce non-explicative variables, plus geographical factors. The income of the census track where the customer lives (Log-Income), the effect of geographical distance between the customer and their favorite supermarket (Log-Distance) and the orders made by the five closest neighbors (Orders-Neighbors) (removing non-significant variables in Baseline model). The results, which are in the second column of Table 3.3, are clear: firstly, individuals living in economically depressed zip codes tend to be less loyal and stop using the application more than in zip codes with higher income. Similar conclusions were found in İlhan and İşçioğlu (2015) and Kee and Wan (2004) who conclude that as level of income increases, the probability of making online grocery shopping increases. This is evidence of the purchasing power of different areas of Madrid and the lifestyle of people living in. So, they did not mind sacrificing price (paying the fee) to be released from dull tasks such as grocery shopping. Additionally, we believe that opportunity cost is something behind this conclusion, since the results show that higher income people find more interesting the application which could help them to spend time in more rewarding tasks. Secondly, the dot-com company is more likely to lose customers if they live close to their favorite supermarket which is reasonable as the customer is not willing to pay the application service cost if the market place is not that far. These results are in the same line of Elms et al. (2016) research, who explains in his case study how the distance and traffic congestions might encourage people to use online shopping services.

The last geographical-related variable to take into account is the orders made by the five closest neighbors of each customer which appear to be a good predictor. So, apart from the data from each customer, the neighbors' information matters to predict future connection between customer and the company. In this case, when a customer is being surrounded by other users that in their first week using the application made quite a few orders, this is an indicator that the customer probably will remain using the application.

Yet again, the generalized I Moran test rejects the null for W_{5nn}. As we note in methodological section, the presence of spatial autocorrelation in the residuals lead to inconsistent and inefficient estimations (McMillen, 1992).

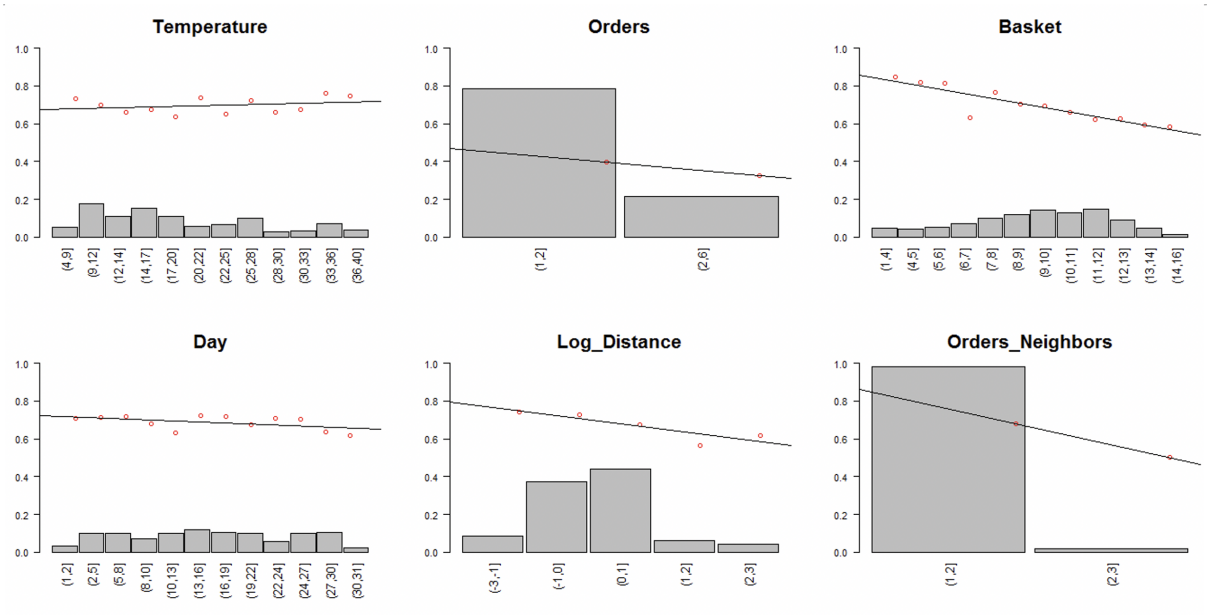


Figure 3.4: Churn rates and histograms (in segments) for the continuous explicative variables

3.3.4 The non-lineal probit model

It is possible that the identification of spatial autocorrelation could be explained by the omission of non-linear relationships between response variable and explicative factors (Basile et al., 2014). With the objective to identify non-linarites the churning rate for each continuous variable split into segments are depicted in Figure 3.4. However, it is barely possible to conclude beyond the positive or negative correlations without a multivariate analysis which can be found in Figure 3.5. A Generalized Additive Model (GAM) (Hastie and Tibshirani, 1986) has been carried out to identify non-lineal relationships. The results represented in Figure 3.5, show that temperature impact linearly in our model from 20 Celsius degrees onwards, basket of products has a direct effect up to 10 products and the log-distance has clearly two slopes to be estimated from minimum to zero and from zero on. Therefore, the Model 3 includes nonlinear patters found in these three variables.

The relevance of the three geographical variables included in the Model 3 can be deduced from the increase in the area under the ROC curve both in the training sample and in the test one. However, the spatial autocorrelation in the residuals is persistent and the Moran tests for spatial independence newly reject the null.

3.3.5 Spatial probit model with non-linearities

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. Previous

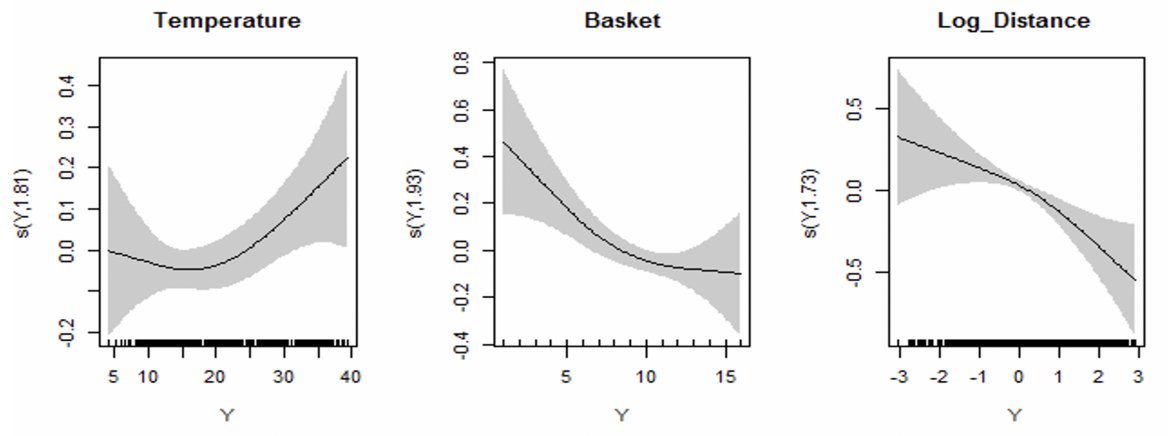


Figure 3.5: GAM multivariate analysis where inflection points were found

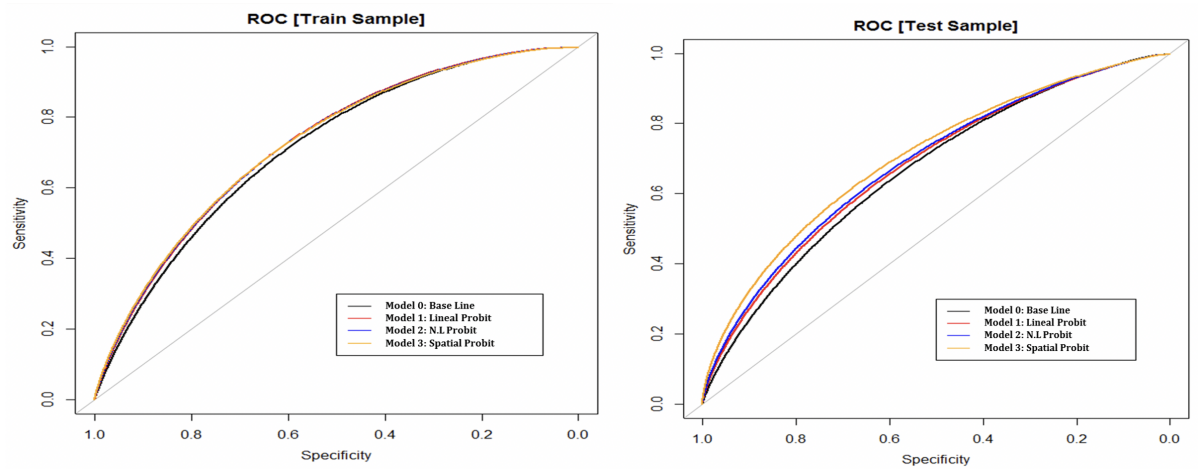


Figure 3.6: Urban area with spatial distribution of individuals

models (Model 0, 1 and 2) have tried to incorporate the spatial effects found by the Join-Count statistics but the presence of spatial autocorrelation in the residual is persistent. The spatial-probit model includes expressly spatial spillover in the specification. The importance of such effects is captured through the new estimated parameter (ρ in Table 3.3) considering the effects of the customer's k = 5 nearest neighbors. The estimated spatial autoregressive coefficient ρ=0.129 is positive and significant.

The better performance of our final spatial-probit model with nonlinearities is reflected in a higher value of the area under the ROC curve. Figure 3.6 shows the improvements in ROC curves both in train and test samples. 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. Also, it proves the more efficient modelling using the spatial probit technique. The stability of the effects of the non-spatial variables denotes a high level of robustness in our results.

3.3.6 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 spatial-probit model. The reason is that, because of the spatial lag of the latent dependent variable WY^*, 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 neighboring individuals. A full description of interpretation of direct, indirect and total effects can be found in Lacombe and LeSage (2015).

Following this methodology, Table 3.4 illustrates the direct and indirect effects of the spatial probit model. First, it is important to highlight that the number of orders made during the first week experience with the company services and the entrance channel by email referred by another user are the most relevant variables to determine churning. The total marginal effects (direct + indirect) of both variables are over 20%. It means that the best commercial strategy distributing the application has been carried out by its own users recommending the application to other potential users. There is obviously no cost for the recommendations and the likelihood to have these customers active more than a week is over 20% higher than other channels. Additionally, creating a good plan for the first impression to the customer is vital. If the company strengthen the connection between the customer and the application during the first week by trying him to make orders, there is over 20% of more chance by each of these orders that the customer remain using the application in the long run.

Secondly, the users coming from Facebook's adverts and those who login through the mobile application are aspects with a meaningful impact on the probability of the customer activity with the e-commerce service. Having a customer coming from Facebook reduces the likelihood of churning in 4% and if the customer started using the service through the mobile application makes it 7% further. This is clearly information which should be used to modulate and optimize investments in marketing actions. Consistent with these conclusions, e-companies might want to control their offers to customers subject to the date as it plays a significant part in defining how loyal to the company the customers would be in future.

Table 3.4: Direct, indirect and total effect of the spatial probit model

	Direct effect	Indirect effect	Total effect
Channel			
Facebook	-0.0352	-0.0051	-0.0403
Email	-0.2413	-0.0349	-0.2762
Login Moment			
Day	-0.0021	-0.0003	-0.0024
App	-0.0676	-0.0098	-0.0774
First Week Experience			
Orders	-0.2012	-0.0291	-0.2303
Geographical Variables			
Log-Income	-0.0389	-0.0056	-0.0446
Orders-Neighbors	-0.0429	-0.0062	-0.0491
Transformed into non-linear			
h(Temperature,25)	0.0071	0.0010	0.0081
n(Basket,10)	-0.0156	-0.0023	-0.0179
h(Log_Distance,0)	-0.0411	-0.0059	-0.0470

Finally, in our research, we noted that the geography plays a key role in the sustainability of a customer portfolio. It is worth using geographic variables in the analysis, as it gives a sense of how dominant the dot-com company would be in certain neighbors. Those customers far from the relevant supermarket will have 4.7% per log-kilometer more continuity using the application. Although, this relationship is not continuous but starts from a log-1km. This information is a good indicator for the online company in order to establish new business areas to operate. In addition, in places where log-income per capita is lower, there is a potential risk for disengagement among the customers. In our sample, an additional point in the Log-Income per capita variable decreases the probability of churning by 4.4%. This likelihood might be decreased even more, if customer lives in an area where high activity indicators of the application are registered. Specifically, additional orders of closest neighbors reduce the churn probability by 4.9%.

3.4 Conclusions and business management implications

E-commerce offers companies an unparalleled capacity to expand and capture new business. From customer point of view, it is a powerful instrument to reduce travel costs and saving time. The application of electronic commerce has been growing over the last years (Frasquet Deltoro et al., 2012). Nevertheless, as shown in Gallego et al. (2016), electronic commerce in Spain shows a low growth where even a great percentage of population has never used this type of services. These indicators are even worse focusing on supermarket e-commerce where fresh and food products are sold.

The data used in this study belongs to one of the top dot-com e-commerce companies in Spain and even though for the given reasons the volume of data might not seem very high, it indeed provides good insights and contributes greatly on customer behavior exploration. Small revenues compared to consolidated traditional grocery distributors do not necessarily mean that it is not worthwhile to study. In fact, these types of analysis should give the understanding to expand these businesses which are in continuous expansion, to know in which type of clients need to focus their commercial strategies and obtain the vision to develop the sales activity.

All these facts make it essential to efficiently manage the relationship with the customer to foment long term relationships, to offer attractive promotions and thereby achieve beat one of the greatest barriers of this type of commerce that is linked to the fear of dissatisfaction and the unknown.

In order to better manage grocery e-companies and have a long-term customer portfolio, companies need to fully understand the effect of the main determinants of churn customer choice. Controlling a positive net-inflow in active customers is something that e-companies need to focus. Special promotions and marketing actions should put into practice to make customers loyal. Mozer et al. (2000) confirm that incentives should be offered to those customers whose probability is above a threshold. Churn customer prediction has been discussed previously in literature; however we think that this methodological improvement presented in this paper by introducing spatial econometrics techniques, will help to gain a better understanding of the problem.

Traditional econometric models assume independence among customers' decisions. This assumption could generate inaccurate estimations of parameters that may have an economic impact on the results. In an urban environment, customer decisions seem unrealistic that are not influenced by the decisions of close people. Those spatial spillovers could be explained by direct interaction between neighbors in customers or by the omission of relevant factors (with spatial structure in the model that could exhibit spatial dependence; LeSage and Pace, 2009b).

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. In this paper we explain the churning event by paying special attention to geographical information of customers of a dot-com e-commerce company that provides its services in the great urban area Madrid (Spain). Reaching to the conclusion that spatial model outperforms the non-spatial in terms of prediction power. Our results provide evidence that the probability of stopping using the application increases if nearby customers also churn due to the spillover effects. Furthermore, the use of spatial information regarding location of shopping centers in the city and also using information on the economic welfare of Madrid's areas, provide interesting conclusions about where to optimize the customer

acquisition. An additional log-kilometer of distance between customer and supermarket reduces the probability that such customer becomes inactive by 4.7%. In addition, if customer lives in an opulent area the probability sharply decreases even more. Hence, spatial distribution of the customer portfolio is a big deal for those startups companies dedicated to e-commerce. Moreover, it is demonstrated that certain actions of neighbors such as the orders requested, improve the model of churning. An additional order by the five closest neighbors in the past, reduces the probability of leaving the company by 4.9%.

Another contribution of the present paper is the use of GAM methodology to reveal behaviors and patters hidden in a linear model. Non-linearity effects are explored so that the model has the best possible performance. Our results indicate that, e-commerce shopping companies should focus on distances between customers and market stores. When distances are very narrow, online company will have to make an extraordinary effort to retain customers and those customers. An additional relevant point is that, from the first kilometer on seems to be the inflection point whereby the churning begins to decrease. Another non linearity found was in temperature the day of signing up. As the atmospheric temperature increases from 25 degrees, customers tend to be less loyal and probably will use the application once. Although our initial expectations were to find correlation both in high and low temperatures (extremes). Finally, our finding reveals that it only affects when temperature is really high.

In this paper, it is also quantified the effect on churning of customers coming from social networks. Direct recommendations and Facebook apparently are vital to capture loyal customers, so investments on advertisement on this platform and giving the customer the possibility to share the application's installation is economically worthwhile in the long run. It is also demonstrated that first experience of e-customers using the application is significant to infer future behavior, so a good offer strategy would lead to a sustainable customer portfolio.

Finally, few areas of further research should be noted. First, this paper is focus on churning in the very first moments of the e-customer. So, as these dot-com businesses are expanding nowadays, more information will be available to understand patterns in future stages. Moreover, as behavior dynamism through time should be tested introducing time level to the spatial regression. Secondly, exploring other territories and compare results to this paper's findings would be interesting to see how estimators change depending on regions or countries. Lastly, in this paper we have demonstrated the importance of distance (in kilometers) between customers and supermarket 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.

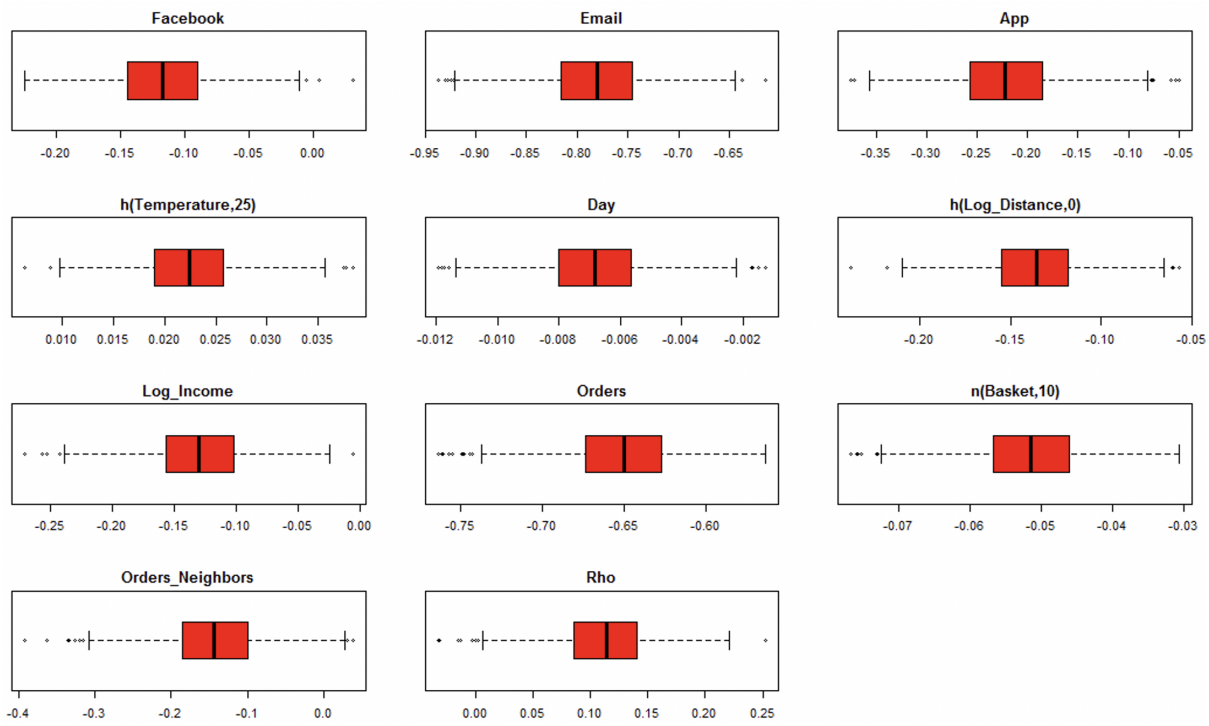


Figure 3.7: Box plot of cross validation estimators per variable

3.5 Appendix

In order to validate the results obtain in Model 3 a cross validation has been made to validate the estimations. Figure 3.7 showed the results to save space. By taking 80% of the training sample randomly for one thousand times, the variability of the estimators in spatial-probit model is obtained. The results confirm the existence of a positive and significance contagious effect, or spillover effect in all the simulations. The significance of the new ρ parameter also means that Model 4 outperforms the previous nested models.

Chapter 4

Searching the correct specification in Spatial Probit Model

4.1 Introduction

There is a well-known aphorism of Box (1980) that, although a little simplistic, reflects a widespread position about how to specify an econometric model: "All models are wrong; some models are useful". In the case of spatial econometrics, the selection of the more adequate model is an open problem and, as far as we know, only a reduced number of papers has been oriented to give some light about this topic.

In the specific case of continuous spatial econometric models some contributions have been trying to find the true data generating process (Florax et al., 2003; Mur and Angulo, 2009; Agiakloglou and Tsimpanos, 2021). Perhaps the Florax et al. (2003) paper is the more relevant contribution (it is almost the most cited paper related with this topic) where the authors use the well know Lagrange Multipliers (LM) tests with the objective of identify the most adequate spatial generation mechanism of an observed dataset. Two strategies were proposed by Florax et al. (2003), the classical 'Specific to general' (Stge) versus the 'General to specific' (Gets) approach or the Hendry approach (Hendry, 1979). The main advantage of use the LM tests is that is not necessary estimate the models under the alternative. This is relevant because in case of medium or big sample size, the estimation by Maximum Likelihood (ML) of a spatial regression model is not a simple task, especially in at the beginning of the 21st century where the power of computers was much lower than the current and the available specific software was limited. The second relevant contribution is the paper of Mur and Angulo (2009), where the authors contemplate slight variations alternatives to Gets/Stge strategies through a Monte Carlo experiment. In this paper, the authors consider four spatial processes, namely SAR, SEM and SARAR and the non-spatial continuous model (spatial independence model, SIM). The results are quite diffuse, in the sense that we do not find conclusive evidence in

favour of either of these two approaches. However, it should be recognized that the Gets strategy seems to be more robust to the existence of anomalies in the Data Generating Process (DGP). Finally, in a recent paper, Agiakloglou and Tsimpanos (2021), develop another Monte Carlo experiment to select the most appropriate spatial regression model considering only the three most frequently applied in practice, such as, SAR, SMA, and SARAR.

In case of spatial econometric models with limited dependent variable, or discrete choice models, the spatial probit model has been the usual instrument to model choices of n individuals under the assumption of spatial interaction. As in the case of continuous spatial econometric models, several tests have been developed to evaluate the presence of spatial correlation in the residuals of a non-spatial probit model (see by example, Amaral et al., 2013), but those tests have been developed under a generic alternative hypothesis and therefore when the null is rejected the researcher don't have information about the true spatial generation mechanism of an observed dataset. Distinguishing between the different type of spillover effect in spatial probit model is essential as there is a serious risk of making incorrect inferences when estimating a misspecified model. Since there are no Monte Carlo experiments, like in case of continuous regression spatial models, to shed any light on the true DGP, different authors of applied papers have been considered different alternatives.

In mostly of cases, the author(s) selected a type of spatial model with the objective of incorporate spatial effects after test the presence of spatial autocorrelation in the residual of a non-spatial probit model using some diagnostic test like Moran test (Amaral et al., 2013) or Join-Count tests (Cliff and Ord, 1981). The SAR probit specification is the most frequently specification and no alternative spatial probit specifications like SDM, SEM or SARAR has been proposed to incorporate spatial effects. The range of scientific fields where the spatial probit models has been applied is enormous, from actuarial sciences De la Llave et al. (2019a), health Ortega-García et al. (2020) or e-commerce De la Llave and López (2020), just to list a few. Less frequently, other papers consider several spatial probit specifications and compare them in different ways. For example, Läpple et al. (2017) estimate SDM, SDEM, SLX and SIM and compares them in terms of predictive power based on percentages correctly predicted outcomes in the sample. Yang and Knook (2021) estimate an spatial Durbin probit model and the authors compare the non-spatial probit model (SIM) with SDM based on the value of the likelihood and the McFadden R^2 coefficient. In this paper the authors use a Wald test to confirm that the SDM probit model should not be reduced to the model that includes one spatial effect, that is the spatial autoregression (SAR) or spatial error (SEM) probit model. Yang and Sharp (2017) compare four spatial probit model based on the highest posterior model probability (SAR; SEM; SDM and SLX) estimated by the Bayesian Markov Chain Monte Carlo. Mate-Sánchez-Val (2021) uses the Likelihood Ratio (LR) test to select the model between SAR

and SEM selecting the specification with highest LR value.

All those evidences show that no information about the correct spatial DGP for spatial probit model is available for the researches that uses this methodology and therefore, it must be necessary to give some light to the correct selection of the model. Without a process that verifies the specification of the spatial probit models, the validity of the Axiom of Correct Specification Leamer and Leamer (1978) cannot be assured. Hence, it could not be guaranteed that there is no correlation between the researcher's beliefs and the final model (Mur and Angulo, 2009). Therefore, the aim of this paper is to contribute further evidence to the debate, outlined briefly above, about how to specify and develop better spatial discrete choice models. At a time when only the investigation of Beron and Vijverberg (2004b) show some evidence using the LR test.

This study conducts a comparison of five spatial probit models (SIM; SAR; SEM; SLX and SDM) through several selection strategies, using an extensive Monte Carlo analysis. Our findings show in the first instance that the misselection of an econometric model has severe consequences. This produces bias the model parameters and therefore it affects the final use of the probit model. This has been an open debate in previous studies carried out with continuous spatial models (LeSage, 2014b; Rüttenauer, 2019). This analysis extends those investigations to the field of spatial probit. In line with previous papers, SDM outperforms the most used models SAR and SEM. Another contribution is the evaluation of a Stge and a Gets strategy for the correct selection of spatial probit models. Since the academic debate on the probit environment has not started, our proposal is the introduction of two selection flows, evaluating their ability to identify true DGPs under ideal and non-ideal data conditions. The results are broken down into various levels to see proportions of success. In addition, the proposed methodology is compared with the result that a Gradient Boosting algorithm would blindly provide. In line with Mur and Angulo (2009), it is difficult to decide between Stge and Gets strategies under ideal conditions. Both strategies behave quite well, specifically when the sample is greater than 900 observations or when the dependency is high. By introducing soft non-ideal conditions such as endogeneity or lack of information in the specification, the strategies slightly lower their performance. On the other hand, when the endogeneity is very severe, the SEM identification becomes complicated.

Section 4.2 reflects the strategies to follow to choose the correct DGP and the tests used for the process. Section 4.3 shows the designed Monte Carlo experiment. Subsection 4.4.1 gives evidence on the consequences on the model coefficients of not correctly selecting the appropriate model. Section 4.4.2 contains the results of the Stge and Gets strategies, as well as their comparison with a GBM algorithm. Finally in section 4.6 the conclusions of the analysis.

4.2 Model selection in a spatial probit context

The discussion focuses on finding the best way to identify the true functional form of the model given some data. The idea in both Gets and Stge is to start with a suitable specification and use different tests to assess its suitability. In principle there is no preference for any strategy. Access to the evaluation tests is free and the estimation of the models is much faster than it was years ago. In any case, for the design of the algorithms, a simplification criterion has been followed (having a reasonable result with few tests).

The detection of true DGP is a very complex task. The main problem is that tests designed to detect spatial dependency in the response variable, like Join Count (JC test) (Cliff, 1973), reject the null when in fact there is dependency on the disturbances or when a spatially lagged variable is present to understand the phenomena. It also happens that the t-test reacts for example detecting significance in spatially delayed independent variables when in fact there is autocorrelation in the lag of the dependent variable or in the error. When analyzing the residuals of a model, we obtain very relevant information on whether the residuals are white noise or whether there is still spatial structure in the disturbances. In Pinkse and Slade (1998), Kelejian and Prucha (2001b) (KP test) and Pinkse (2004), there are three tests proposed focused on finding patterns in the residuals of the spatial probit. All of them work well detecting SAR, SEM, SDM structures within SIM residues, especially Kelejian and Prucha (2001b) according to Amaral et al. (2013). But all of them leave open the type of spatial pattern that the residuals hide. Furthermore, when applying these tests to the residuals of spatial specifications, we are unable to distinguish those that eliminate completely the spatial dependency. Agiakloglou and Tsimpanos (2021) performs Monte Carlo analysis via the Likelihood Ration (LR). LR is the most widely used test to compare spatial probit models. This test, as shown in Agiakloglou and Tsimpanos (2021), manages to correctly select 70%[1] of the cases only by deciding between SIM, SAR and SEM. Finally, The most widespread methods for comparing models are the Akaike Information Criteria (Akaike, 1998) and the Bayesian Information Criteria (Schwarz, 1978). These methods provide simplicity and speed when choosing between models. At the moment there is no evidence on the power for discriminating spatial dichotomic models. However, on continuous spatial regressions, Agiakloglou and Tsimpanos (2021) performs simulations, reaching accuracy levels close to 80% for SEM, SAR and SARAR. All these tests have been taken into consideration to formulate selection strategies Stge and Gets for spatial probit DGPs.

The main difference between both strategies as originally formulated by Florax et al. (2003) is the starting point for the identification of the spatial structure. The broader specification would cover all types of spatial dependency.

[1]For sample sizes of 200 observations, parameter of autocorrelation equals to 0.5, the LR test obtains a 75% of precission for SAR and 65% for SEM

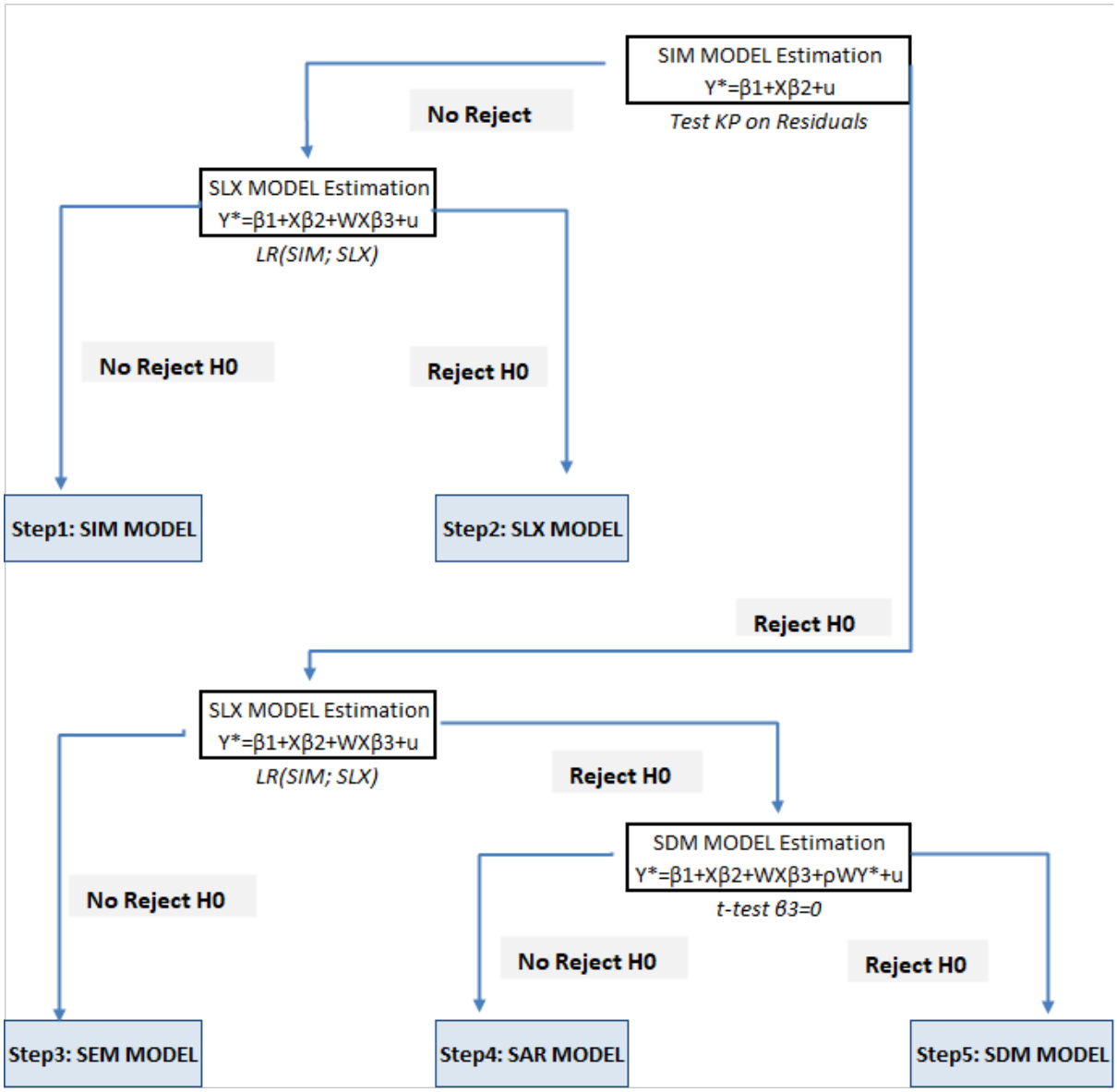


Figure 4.1: Flow for Specific to General (Stge) approach

$$\begin{aligned} Y^* &= \beta_1 + X\beta_2 + WX\beta_3 + \rho WY^* + u \\ u &= \lambda Wu + e \end{aligned} \tag{4.1}$$

When all the spatial parameters β_3, ρ and λ lose their significance, then we are talking about a SIM model, which is the origin of the Stge strategy in which there is no spatial dependency between observations. When λ=0 is not rejected, then we speak of a SDM model that combines spatial dependence on the dependent variable and spatial lag on the exogenous variable. All other resulting models will be SLX when only β_3 is significant. Finally, SAR and SEM in which ρ and λ will be significant respectively.

Figure 4.1 depicts the strategy Stge for probit model selection. The sequence begins

with the analysis of the residuals of a SIM model. If Generalized IMoran test by Kelejian and Prucha (2001b) is not rejected[2], the possible DGP could only be SIM or SLX. In order to distinguish them, we apply the LR test. On the other hand, if the SIM residuals show a significant spatial pattern, then the next step is to estimate the SLX model. Again, we compare through LR the SIM and SLX models. When we do not reject null in favor of SIM, then we decide the estimated DGM as SEM. Finally, we estimate the SDM model in which we will discriminate between SAR and SDM based on the significance of the parameter β_3. In case the parameter is significant, then the estimated DGP will be SDM, otherwise SAR.

The Gets sequence is a bit more complex. It follows the flow in line with Mur and Angulo (2009), although with slight changes, since in this analysis we incorporate the SLX model. The flow is shown in Figure 4.2. The simplification process starts from the SDM model. In the case of not rejecting the hypothesis of β_3=0, we will have two possible outcomes: SIM or SAR. Discrimination between them will be achieved by testing the significance of ρ. In the case of being significant, the process will follow a SAR, otherwise a SIM. When we reject that β_3=0 then we have the full range of models (except SAR) likely to be the estimated DGP. Here we must resort to the LR test between SEM and SDM. This is the so-called likelihood ratio of common factors (LRCOM)[3] appropriate for testing the validity of the SEM (Davidson et al., 2004). By not rejecting the LR, and if definitely the parameter λ of the SEM model is significant, then the final proposal will be SEM, otherwise SIM. Rejecting the LR between SEM and SDM will depend on the significance of the ρ parameter. If ρ is not significant, the process ends pointing to an SLX, and if it is significant then this means that all the parameters of the model are significant, so the original SDM model will be the valid one.

4.3 The design of the Monte Carlo study

This section describes the Monte Carlo process followed to evaluate both the backward and forward methodology. We start from the five different model specifications presented before (SIM, SEM, SLX, SAR and SDM models). The main objective of the study is to compare the model selection methodologies, by evaluating the degree of recognition of true DGP.

The first part of the analysis has been carried out under ideal conditions. This means that there is no endogeneity in the model and all the variables that determine the dependent variable are present in the specification. Furthermore, the simulated residuals

[2]For all the comparisons both in the Stge and the Gets approach, the Bonferroni correction for nested contrasts has been applied in the definition of the p-values. However, results do not chance much if we had applied cut-off p-values.

[3]The Comfac LR degrees of freedom are the model parameters

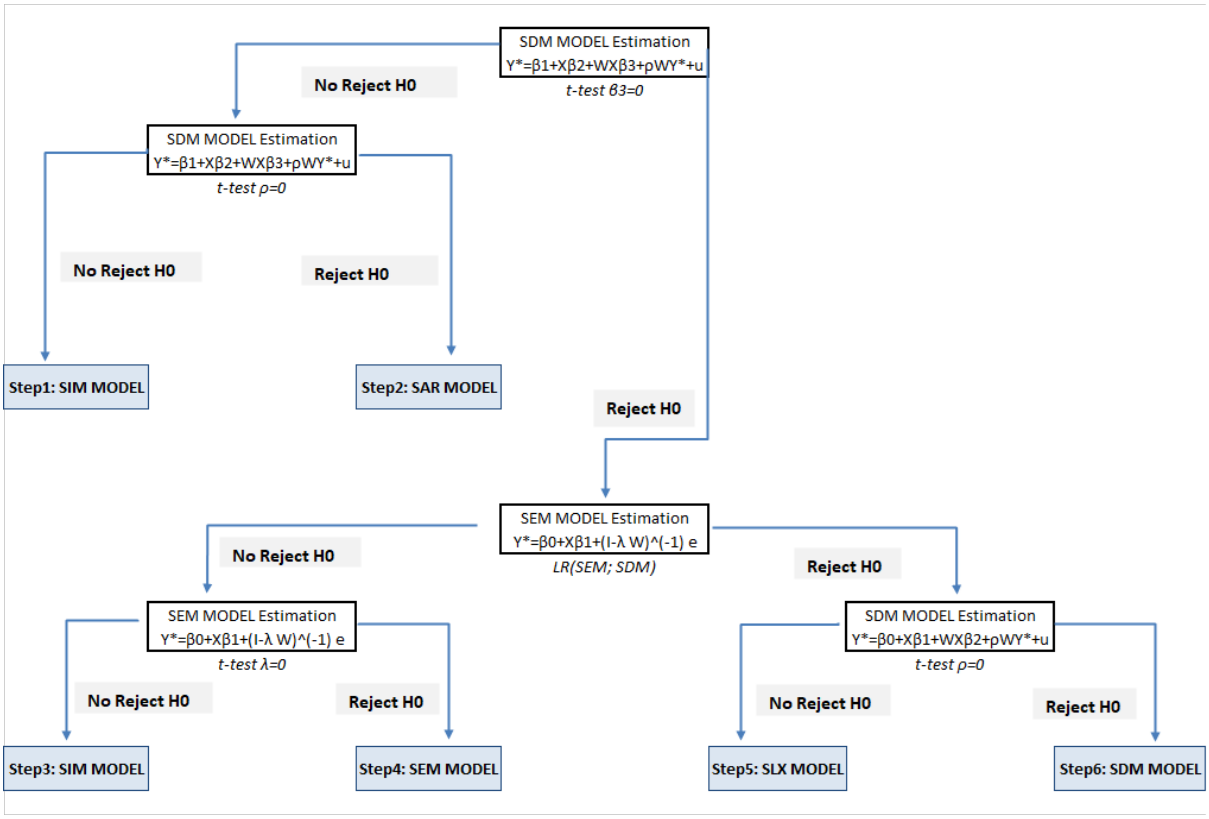


Figure 4.2: Flow for General to Specific (Gets) approach

are distributed as a normal with mean equal to zero and variance equal to one. In a second phase of the study, three different conditions have been introduced in the model, two types of endogeneity have been incorporated at different levels of correlation [0.2% and 0.9%] to see the performance of the selection methods. The third non ideal condition contains simulations where the model specification does not contain a spatially lagged variable which is significant for the model[4]

For our Monte Carlo study, we build the unobservable latent variable following the model defined in 4.1. The response variable Y, object of the modeling, will be the observable part of $Y*$, which will take the value 1 when $Y*$ is positive and 0 when Y* is negative. The parameter values selected for the model have been previously studied in such a way that appropriate characteristics are fulfilled for the formulation of Monte-Carlo's work. The coefficients chosen are β_1 =1; β_2=-0.5 and β_3=-0.4. β_1 refers to the intercept applicable for all models. β_2 to the exogenous variable (X) randomly generated following a Normal(μ=1,σ=2) applicable for all models too. And β_3 to the spatially lagged exogenous variable (X) applicable to SLX and SDM models. These coefficients guarantee a quite reasonable balancing of the dichotomous variable[5]. Furthermore, these coefficients keep the Area Under the ROC Curve (AUC) of all models always greater than

[4]The variable simulated and omitted in the specification is W * V which is randomly generated following a Normal(μ=1,σ=2). The coefficient in the true specification is 0.2

[5]The resulting percentage of Y=1 is between 0.4 and 0.8

0.78.

The number of observations (n) is defined as an element to be studied, so the sample sizes 100, 400, 900, 1600, 2500 are proposed for the simulations. A regular grid division of space has been considered ($\sqrt{n} \times \sqrt{n}$). The real adjacency matrix W is assumed to be known and follows the rook criteria. Then the W matrix is row-standardized. Given the models to be evaluated, we have three types of parameter combinations (ρ,λ). The first combination (ρ=0,λ=0) for SIM and SLX models. The second combination (ρ=ι,λ=0) for SAR and SDM. And the third combination (ρ=0,λ=ι) for the SEM model. For each case, 5 values for ι have been simulated (from 0.3 to 0.7 by adding 0.1).

Each combination has been repeated 500 times. One of the main obstacles of the analysis is obviously the execution time. The simulations have been carried out in R language on the Microsoft Azure cloud platform using a Standard DS3 processor with 14 GB of RAM and 4 Cores. A full run with all combinations takes around 3.5 hours. For estimation of spatial-probit model we use the procedure based on Conditional Maximum Likelihood developed by Martinetti and Geniaux (2017). The ProbitSpatial R-package (Martinetti and Geniaux, 2021) was used to estimate the model.

4.4 Results of the Monte Carlo study

4.4.1 Consequences of an incorrect model choice

The first question to ask is the consequences of an incorrect model choice. Very often, econometric models are used both to explain past phenomena and to evaluate the effect of certain actions on a socioeconomic variable. To cite a few examples, in LeSage et al. (2011), he makes an evaluation of the government aids programs after Hurricane Katrina. In Ortega-García et al. (2017), the factors that affect childhood cancer are sought. In Storm et al. (2015), it is analyzed how much direct payments affect farm survival. This reveals that we must look for parameters that are as accurate as possible, otherwise it will have consequences in the actions that are taken based on the data. Therefore, the analysis of the data and the specification of the model must have a careful consideration. The incorrect selection of the model might cause poor estimators not fully efficient producing biases and when there is an obvious spatial lag, the estimators can be inconsistent. This might cause a misidentification of the real factors involved in the modelling.

The purpose of this section is to show the deviation in the estimation of the parameters when we model with a specification other than the true DGP. The entire simulation process leaves us with a huge amount of data for analysis. Tables 4.1 and 4.2 show average bias in parameters estimated, accuracy through ROC curve Area and their standard

deviations for each. For the sake of brevity, only the Figures for sample sizes of 400 and 2500 are shown. ρ and λ in the table take the value equal to 0.6.

Our simulations indicate that by selecting the correct DGP, the estimates for all cases become very precise and consistent. The problems come up when the selection is incorrect. The first conclusion when looking at the data, is confirmation of the statement in McMillen (1992), "Spatial autocorrelation reduces efficiency and can make OLS parameter estimates inconsistent". When there is a certain spatial pattern in the real data, both β_1 and β_2 in the SIM model present a relevant bias. In addition, it is noticeable the decrease in the ROC in these models. On the contrary, SDM presents the best results in terms of accuracy. SDM has also unbiased β_1 and β_2 estimators specially when sample size is 2500. When real DGP is a SLX, SDM nails the coefficients without giving value to the autoregressive parameter. In case of a DGP=SAR, the bias in ρ is positive but quite small and compensates with β_3 which has a negative bias. In case of DGP=SEM, SDM has a significant bias in β_1 and solves the spatial problem giving value to β_3 and ρ.

SAR and SEM models, which actually are the most used models in research, present bias when the type of spatial dependency contained in the real data is incorrect. When real DGP follows a SDM then SAR and SEM have relevant bias in a very similar way each other. They tend to overestimate the dependency parameter and both β_1 and β_2 are underestimated. Finally, when the true DGP is an SLX, SAR and SEM are biased in different directions at β_2, however, just SAR modelling gives value to the spatial dependency parameter.

Regarding modelling with SLX, it obtains a good performance in terms of accuracy. The estimator β_2 is quite centered for all DGPs. When DGP is SAR or SDM, β_3 tends to take on a relevant value, in turn decompensating the true value of β_1. However, in case of a DGP=SEM β_3 it remains unchanged.

All these impacts have some modifications when the sample size changes or when the spatial parameter varies. For further analysis on bias in β_1 and β_2, Figures 4.3 and 4.4 have been generated. They represent the bias of each of the estimators given four types of sample sizes and a small (0.3) and large (0.6) autocorrelation both in ρ and in λ. It can be seen how for β_1 the bias remains constant for different sample sizes and obviously varies depending on the autoregressive parameter. β_2 tends to have a constant bias for each model not equal to the true DGP when sample size varies. However, when estimating with SDM, it fails to fit the coefficient perfectly when the DGP follows a SAR or a SEM and the sample size is high as well as the special dependency parameter. Again, it is clear that SLX provides good estimates for β_2 when DGPs equal to SAR or SEM, especially if ρ and λ are low. Furthermore, when we find a DGP SDM, SLX is the one that would make the best estimate of β_2 among the compared models.

In conclusion, the analysis of the bias of the estimators and the precision of the models shows SDM as the model in which, without prior knowledge of the type of spatial struc-

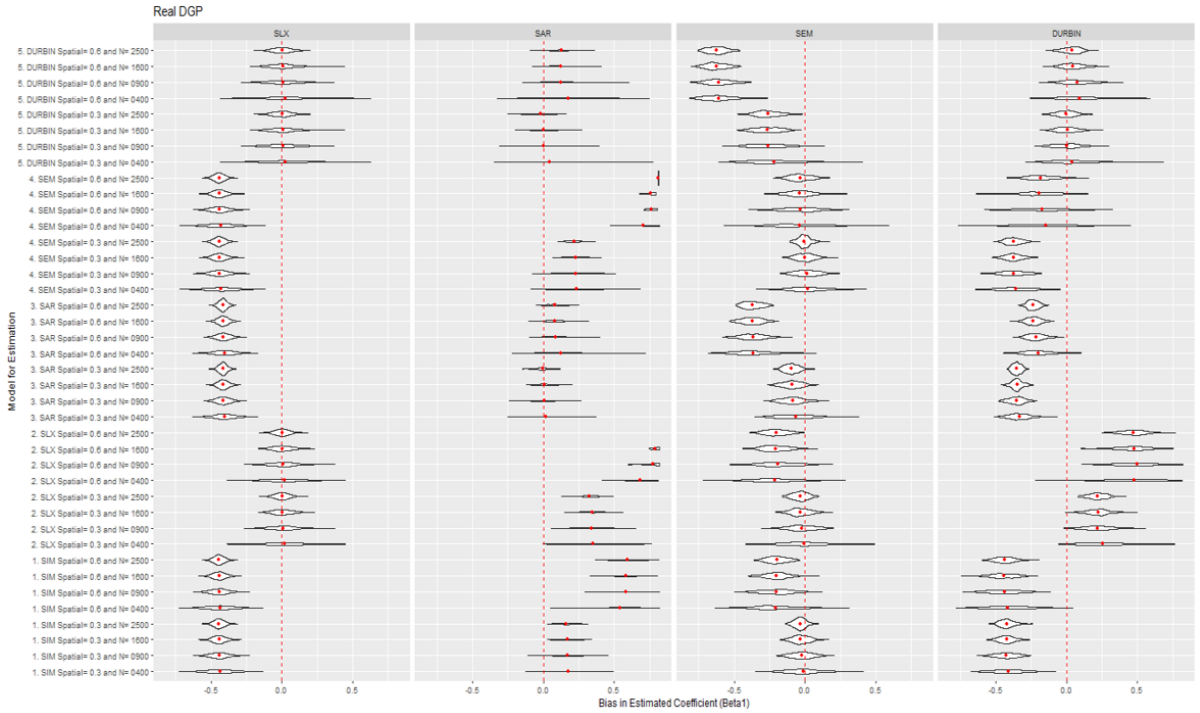


Figure 4.3: Bias in β_1 for different sample sizes and parameters ρ and λ

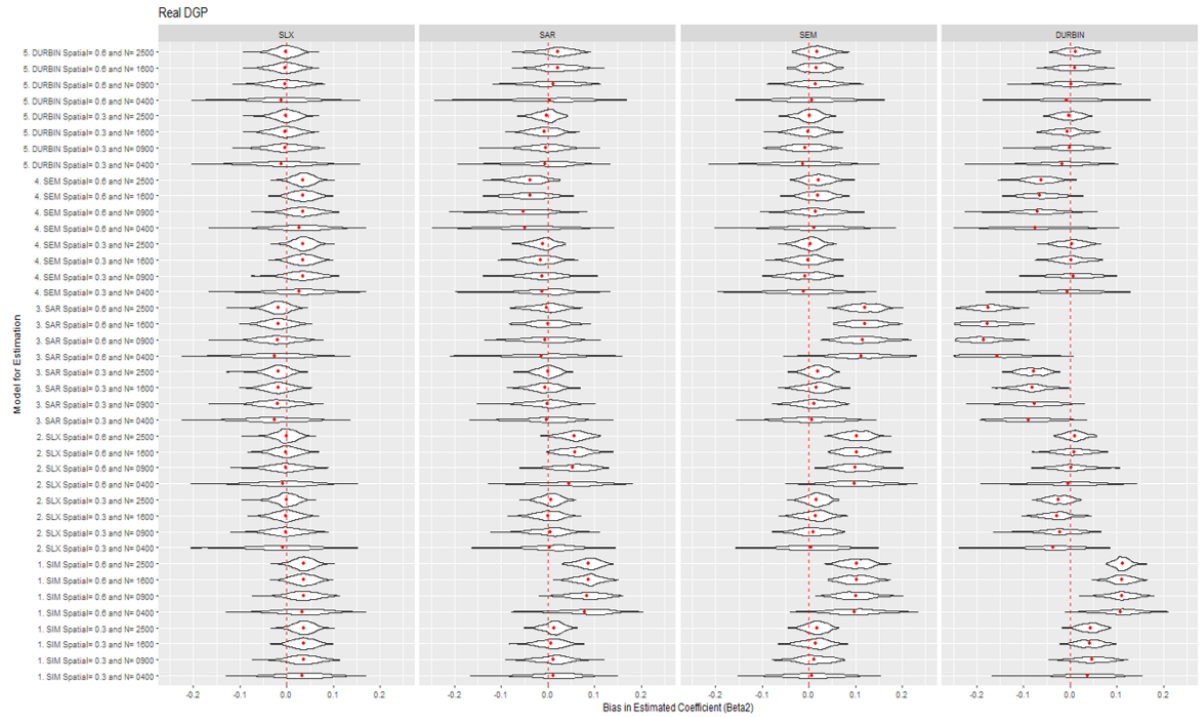


Figure 4.4: Bias in β_2 for different sample sizes and parameters ρ and λ

ture, it obtains better results in terms of sensitivity and specificity. Also, if the sample size is high, the estimators are centered. The main drawback is the interpretability of this SDM model in terms of direct and indirect effects, which is more complex. It has also been seen that SLX can be good. It is true that it produces biased estimators, especially when the parameter ρ is high, but the main advantage is its simplicity of interpretation and calculation.

Table 4.1: Bias, ROC and Standard deviation in brackets when estimating using different specifications under ideal conditions. N=400

Model	Bias β_1=1	Bias β_2=-0.5	Bias β_3=-0.4	Bias ρ=0.6	Bias λ=0.6	ROC
			DGP: SIM			
SIM	0.01 (0.10)	0.00 (0.05)				0.84 (0.02)
SLX	0.01 (0.13)	-0.01 (0.05)	0.00 (0.08)			0.84 (0.02)
SAR	0.01 (0.12)	-0.01 (0.05)		-0.01 (0.11)		0.84 (0.02)
SEM	0.02 (0.10)	-0.01 (0.05)			-0.01 (0.16)	0.84 (0.02)
SDM	0.03 (0.22)	-0.01 (0.05)	-0.01 (0.11)	-0.02 (0.16)		0.84 (0.02)
			DGP: SLX			
SIM	-0.43 (0.09)	0.03 (0.05)				0.82 (0.02)
SLX	0.02 (0.13)	-0.01 (0.05)	-0.01 (0.08)			0.85 (0.02)
SAR	-0.40 (0.08)	-0.03 (0.06)		0.37 (0.08)		0.84 (0.02)
SEM	-0.43 (0.10)	0.03 (0.05)			0.01 (0.17)	0.82 (0.02)
SDM	0.03 (0.17)	-0.01 (0.06)	-0.01 (0.10)	0.00 (0.14)		0.85 (0.02)
			DGP: SAR			
SIM	0.60 (0.22)	0.08 (0.05)				0.82 (0.03)
SLX	1.11 (0.27)	0.04 (0.06)	-0.38 (0.10)			0.85 (0.03)
SAR	0.12 (0.16)	-0.02 (0.08)		0.06 (0.08)		0.86 (0.03)
SEM	1.11 (0.33)	-0.06 (0.09)			0.71 (0.10)	0.82 (0.03)
SDM	0.17 (0.21)	0.00 (0.07)	-0.06 (0.11)	0.04 (0.09)		0.86 (0.03)
			DGP: SEM			
SIM	-0.21 (0.17)	0.10 (0.05)				0.79 (0.02)
SLX	-0.21 (0.18)	0.10 (0.05)	0.01 (0.09)			0.79 (0.02)
SAR	-0.37 (0.14)	0.11 (0.05)		0.41 (0.11)		0.79 (0.03)
SEM	-0.04 (0.20)	0.01 (0.06)			0.02 (0.09)	0.79 (0.02)
SDM	-0.62 (0.12)	0.01 (0.06)	0.31 (0.11)	0.61 (0.08)		0.80 (0.02)
			DGP: SDM			
SIM	-0.42 (0.32)	0.11 (0.08)				0.79 (0.04)
SLX	0.51 (0.45)	0.00 (0.12)	-0.37 (0.21)			0.87 (0.04)
SAR	-0.20 (0.22)	-0.20 (0.18)		0.24 (0.10)		0.89 (0.05)
SEM	-0.15 (0.47)	-0.08 (0.15)			0.82 (0.14)	0.79 (0.04)
SDM	0.09 (0.33)	-0.01 (0.14)	-0.07 (0.22)	0.08 (0.15)		0.89 (0.04)

Table 4.2: Bias, ROC and Standard deviation in brackets when estimating using different specifications under ideal conditions. N=2500

Model	Bias β_1=1	Bias β_2=-0.5	Bias β_3=-0.4	Bias ρ=0.6	Bias λ=0.6	ROC
			DGP: SIM			
SIM	0.00 (0.04)	0.00 (0.02)				0.84 (0.01)
SLX	0.00 (0.05)	0.00 (0.02)	0.00 (0.03)			0.84 (0.01)
SAR	0.00 (0.05)	0.00 (0.02)		0.00 (0.04)		0.84 (0.01)
SEM	0.00 (0.04)	0.00 (0.02)			0.00 (0.07)	0.84 (0.01)
SDM	0.00 (0.08)	0.00 (0.02)	0.00 (0.04)	0.00 (0.07)		0.84 (0.01)
			DGP: SLX			
SIM	-0.44 (0.04)	0.04 (0.02)				0.82 (0.01)
SLX	0.00 (0.05)	0.00 (0.02)	0.00 (0.03)			0.85 (0.01)
SAR	-0.42 (0.03)	-0.02 (0.02)		0.37 (0.04)		0.84 (0.01)
SEM	-0.44 (0.04)	0.04 (0.02)			0.00 (0.07)	0.82 (0.01)
SDM	0.00 (0.07)	0.00 (0.02)	0.00 (0.04)	0.00 (0.06)		0.85 (0.01)
			DGP: SAR			
SIM	0.59 (0.09)	0.09 (0.02)				0.82 (0.01)
SLX	1.09 (0.11)	0.06 (0.03)	-0.38 (0.04)			0.85 (0.01)
SAR	0.08 (0.06)	0.00 (0.03)		0.07 (0.03)		0.86 (0.01)
SEM	1.08 (0.11)	-0.04 (0.03)			0.72 (0.04)	0.82 (0.01)
SDM	0.13 (0.08)	0.02 (0.03)	-0.06 (0.05)	0.06 (0.04)		0.86 (0.01)
			DGP: SEM			
SIM	-0.20 (0.07)	0.10 (0.03)				0.79 (0.01)
SLX	-0.20 (0.08)	0.10 (0.03)	0.00 (0.03)			0.79 (0.01)
SAR	-0.37 (0.06)	0.12 (0.03)		0.43 (0.05)		0.78 (0.02)
SEM	-0.03 (0.07)	0.02 (0.03)			0.03 (0.05)	0.79 (0.01)
SDM	-0.62 (0.06)	0.02 (0.02)	0.30 (0.05)	0.62 (0.05)		0.79 (0.01)
			DGP: SDM			
SIM	-0.44 (0.13)	0.11 (0.03)				0.79 (0.02)
SLX	0.47 (0.18)	0.01 (0.04)	-0.36 (0.09)			0.87 (0.02)
SAR	-0.24 (0.08)	-0.18 (0.06)		0.24 (0.05)		0.88 (0.02)
SEM	-0.18 (0.19)	-0.06 (0.06)			0.83 (0.05)	0.79 (0.02)
SDM	0.04 (0.12)	0.01 (0.05)	-0.04 (0.10)	0.09 (0.06)		0.89 (0.02)

4.4.2 Results of the selection strategies

The question to be resolved in this subsection is to know how much we can trust these two strategies given different casuistries of the data. Table 4.3 summarizes the percentage of simulations that correctly select the true DGP under ideal conditions. For both strategies, there is a notable growth in the percentage of correct answers as the sample size increases. From a size of 900 observations the performance is quite regular and reasonable. The same occurs with the spatial dependency level. In the case of SIM and SLX there is no spatial dependence within the model, but in the case of SAR, SEM and SDM there is a more or less constant increase in the probability of success when the autocorrelation

parameter increases.

Under ideal conditions, both strategies have a very similar level of performance. Only Stge seems to outperform Gets in SEM identification. Both algorithms reach levels close to 1 when the sample size is high (> 400) and when ρ or λ show high level of spatial dependency. Actually, the whole average of success is 97.08% for Gets and 97.46% for Stge. SDM is the real DGP which is detected with almost 100% of accuracy. On the other hand, for problems where there are not many records (< 400) and the spatial dependence is low (< 0.4), it is difficult to select correctly the true DGP. The precision for Gets is 39.69% and for Stge is 41.84%, being SEM the model that is most difficult to be identify for Gets with 12% of probability of success and SAR for Stge strategy with 20%. Finally, if comparing both strategies when n=900 with ρ and λ either 0 or 0.5, then Gets shows and average of 91.63% and Stge 93.95%.

A point that also deserves special attention is when the algorithm fails. When the algorithm goes wrong, given the structure of the strategies and the combinations of tests, they make the strategy point to another specific model, instead of being distributed evenly among the rest of the models. This can be seen in 4.5. For example, both strategies select 5% true SEM as SIM. However, Gets confuses true SEM much more with SAR (8% instead of 2% Stge). SDM is the last Step in both strategies to be tested. Everything that has not been selected already falls into SDM category. It works for both strategies quite well and just 8% of the cases are selected as SDM when in reality they are not.

Table 4.3: Percentages of correct identification of the DGP under ideal conditions

			n=100	n=400	n=900	n=1600	n=2500
DGP	ρ	λ	Stge (Gets)	Stge (Gets)	Stge (Gets)	Stge (Gets)	Stge (Gets)
SIM	0.0	0.0	0.94 (0.98)	0.94 (0.97)	0.94 (0.97)	0.93 (0.96)	0.95 (0.97)
SLX	0.0	0.0	0.57 (0.31)	0.98 (0.92)	0.98 (0.97)	0.97 (0.96)	0.98 (0.97)
SAR	0.3	0.0	0.06 (0.08)	0.23 (0.36)	0.59 (0.77)	0.91 (0.95)	0.97 (0.97)
	0.4	0.0	0.05 (0.16)	0.46 (0.66)	0.88 (0.92)	0.99 (0.99)	0.98 (0.98)
	0.5	0.0	0.14 (0.29)	0.71 (0.87)	0.98 (0.98)	0.97 (0.97)	0.98 (0.98)
	0.6	0.0	0.22 (0.35)	0.87 (0.94)	0.97 (0.97)	0.96 (0.96)	0.94 (0.94)
	0.7	0.0	0.33 (0.53)	0.95 (0.99)	0.99 (0.99)	0.93 (0.93)	0.93 (0.93)
SEM	0.0	0.3	0.07 (0.03)	0.29 (0.11)	0.57 (0.39)	0.77 (0.58)	0.87 (0.73)
	0.0	0.4	0.14 (0.04)	0.54 (0.31)	0.80 (0.66)	0.95 (0.86)	0.97 (0.96)
	0.0	0.5	0.24 (0.06)	0.77 (0.44)	0.98 (0.84)	0.97 (0.95)	0.98 (0.99)
	0.0	0.6	0.42 (0.16)	0.93 (0.71)	0.97 (0.93)	0.98 (0.98)	0.98 (0.98)
	0.0	0.7	0.64 (0.25)	0.95 (0.89)	0.97 (0.97)	0.98 (0.98)	0.97 (0.95)
SDM	0.3	0.0	0.03 (0.03)	0.52 (0.51)	0.91 (0.88)	1.00 (0.99)	1.00 (1.00)
	0.4	0.0	0.04 (0.04)	0.83 (0.83)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	0.5	0.0	0.14 (0.14)	0.88 (0.89)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	0.6	0.0	0.16 (0.19)	0.90 (0.90)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	0.7	0.0	0.23 (0.24)	0.82 (0.82)	0.98 (0.98)	1.00 (1.00)	1.00 (1.00)

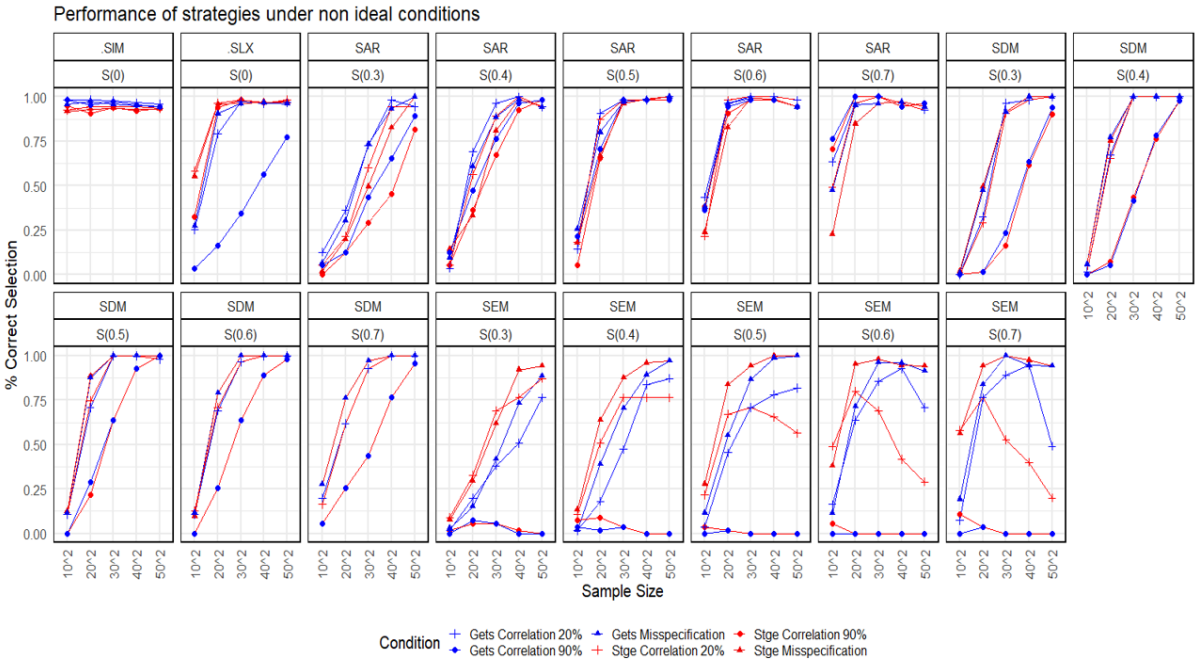


Figure 4.5: Percentages of correct identification of the DGP under non ideal conditions

Ideal conditions hardly exist in real life. For this reason, simulations have been carried out in the presence of real data problems. The first fairly common circumstance that can occur to the analyst is not having all the information that determines a problem. The lack of any significant variable causes a direct impact included in the disturbances of the model. The omission of a relevant variable in the model specification produces a slight decrease in the performance of the strategies. Figure 4.5 shows the degree of performance against sample size and breaks down the graphs by degree of spatial dependence. The decrease in performance is appreciated in the presence of misspecification, which mainly affects the DGP=SAR, especially when ρ is less than 0.5 and also when sample size is lower than 900 observations. The only DGP not impacted is when it follows a SDM which means that all parameters remain significant (Gets Strategy) and KP's null hypothesis for SIM residuals is rejected and LR(SIM,SEM) is also rejected.

Another common problem is endogeneity. Figure 4.5 also shows the results of the Gets and Stge strategies for some simulations in which a correlation between the error and the exogenous variable has been introduced. The dark green and blue contain the correlation at 20%. A general decrease in the probability of success of the strategies is observed. Especially in the case of SEM where it is more prejudicial. These results are totally in line with that mentioned by Mur and Angulo (2009). Generally speaking, Gets resists better the endogeneity in the model, especially when the spatial dependency is high. However, when the correlation is very high (90%), the Gets strategy decreases sharply its success level in the case of the DGP=SLX and both strategies fail to identify the SEM. In this instance, SEM is confused in favor of SIM.

4.5 Play the machine: Gradient Boosting versus Stge and Gets

The last point to cover is about whether with all this bunch of tests at our disposal, we have chosen the best way to select the true DGP. All the decisions made for Stge and Gets strategies make econometric sense to us. The tests used for each of the decisions refer in some way to the DGP we are deciding on. However, we cannot guarantee a global optimum as we have not tried all possible test combinations to see an optimal diagnosis. Our intention now is to make a comparison between the strategies and a Gradient Boosting Model (GBM). GBM is a powerful decision tree ensemble technique, prepared to obtain very precise classifications. Being a non-interpretable technique, we will only use it as a benchmark of successes rather than the strategy that has been followed. By allowing the GBM algorithm to optimize the multinomial problem with a database containing all available tests, we are going to check if there are more powerful alternatives. Clearly, GBM does not pay attention to whether the tests used in each decision make econometric sense, but instead will seek to minimize the function cost. We consider 60% training database and 40% test data to avoid overfitting. The gbm package of R has been used (Greenwell et al., 2020). The hyperparameters set are n.trees=100, interaction.depth=1, n.minobsinnode=10 and shrinkage=0.1. The results of the GBM in 4.4 do not improve much those of the Stge and Gets strategies. GBM improves the results of Stge under ideal conditions by 4.7%. Breaking down this percentage, we see how it underperforms when identifying the independent model SIM. However, it has a better performance in the cases of SAR and SEM when the sample size is less than 400. It is difficult to really know all the crosses of variables and tests that GBM is doing behind. We do not even know if these computed decisions make econometric sense. What is true is that as the sample grows there is no evidence that the GBM offers better results than Stge or Gets strategies. Speaking of non-ideal conditions (endogeneity in the model), we see same effects in GBM results as those produced in the aforementioned selection strategies presented. We believe that this comparison between GBM and strategies, is a validation method to be aware of how close we are to an optimum. We can ensure that having such battery of available tests, we could not have reached a much higher result. There is no doubt that by increasing the number of tests, for example, the Vuong and Clarke tests cited in Mur and Angulo (2009), the Q statistic (Ruiz et al., 2010) or the binary scan test (Kulldorff, 1997), the result is likely to improve.

4.6 Conclusions

The search for the correct specification is one of the topics that has received less attention within spatial econometrics. Over the last 40 years, different taxonomies have emerged

Table 4.4: Percentage of correctly identified DGP using GBM algorithm. Comparition GBM with Gets and Stge strategies

	ρ	λ	n=100	n=400	n=900	n=1600	n=2500
				DGP SIM			
GBM	0.0	0.0	0.90 (0.89) (0.89) (0.86)	0.91 (0.92) (0.86) (0.90)	0.91 (0.91) (0.89) (0.89)	0.90 (0.93) (0.90) (0.90)	0.93 (0.90) (0.92) (0.92)
dif-Stge			-0.04 (-0.03) (-0.06) (-0.06)	-0.03 (-0.03) (-0.05) (-0.03)	-0.03 (-0.04) (-0.05) (-0.05)	-0.03 (-0.02) (-0.02) (-0.03)	-0.02 (-0.03) (-0.01) (-0.01)
dif-Gets			-0.08 (-0.09) (-0.10) (-0.10)	-0.06 (-0.07) (-0.09) (-0.08)	-0.06 (-0.07) (-0.08) (-0.07)	-0.06 (-0.04) (-0.06) (-0.05)	-0.04 (-0.06) (-0.03) (-0.02)
				DGP SLX			
GBM	0.0	0.0	0.43 (0.69) (0.47) (0.63)	0.92 (0.92) (0.91) (0.93)	0.96 (0.93) (0.96) (0.92)	0.94 (0.92) (0.94) (0.93)	0.95 (0.95) (0.95) (0.93)
dif-Stge			-0.14 (0.11) (0.14) (0.08)	-0.06 (-0.04) (-0.03) (-0.03)	-0.02 (-0.05) (-0.03) (-0.04)	-0.03 (-0.04) (-0.03) (-0.04)	-0.03 (-0.04) (-0.02) (-0.04)
dif-Gets			0.12 (0.44) (0.44) (0.36)	0.00 (0.13) (0.75) (0.02)	-0.01 (-0.04) (0.61) (-0.04)	-0.02 (-0.04) (0.38) (-0.04)	-0.02 (-0.02) (0.18) (-0.03)
				DGP SAR			
GBM	0.3	0.0	0.20 (0.13) (0.09) (0.05)	0.62 (0.38) (0.25) (0.44)	0.89 (0.73) (0.49) (0.77)	0.98 (0.98) (0.60) (0.92)	0.97 (0.95) (0.85) (1.00)
dif-Stge			0.14 (0.08) (0.09) (0.03)	0.39 (0.16) (0.12) (0.24)	0.30 (0.13) (0.20) (0.27)	0.07 (0.03) (0.15) (0.09)	0.00 (0.00) (0.03) (0.00)
dif-Gets			0.12 (0.00) (0.04) (-0.02)	0.26 (0.02) (0.12) (0.14)	0.12 (0.00) (0.05) (0.04)	0.03 (0.00) (-0.05) (-0.01)	0.00 (0.00) (-0.04) (0.00)
GBM	0.5	0.0	0.57 (0.31) (0.20) (0.41)	0.91 (0.95) (0.73) (0.90)	0.99 (0.98) (0.96) (0.98)	0.97 (0.98) (0.98) (0.97)	0.98 (0.98) (0.98) (1.00)
dif-Stge			0.43 (0.13) (0.15) (0.23)	0.20 (0.08) (0.08) (0.23)	0.01 (0.00) (-0.02) (0.02)	0.00 (0.00) (0.00) (-0.02)	0.00 (-0.02) (0.00) (0.00)
dif-Gets			0.28 (0.16) (-0.02) (0.15)	0.04 (0.04) (0.02) (0.10)	0.01 (0.00) (-0.02) (0.01)	0.00 (0.00) (0.00) (-0.02)	0.00 (-0.02) (0.00) (0.00)
GBM	0.7	0.0	0.71 (0.73) (0.78) (0.63)	0.98 (0.96) (1.00) (0.94)	0.99 (1.00) (1.00) (0.96)	0.93 (0.96) (0.95) (0.95)	0.92 (0.89) (0.95) (0.94)
dif-Stge			0.38 (0.24) (0.07) (0.40)	0.03 (0.00) (0.00) (0.09)	0.00 (0.00) (0.00) (0.00)	0.00 (0.00) (0.00) (-0.02)	-0.01 (-0.04) (-0.01) (0.00)
dif-Gets			0.18 (0.09) (0.02) (0.15)	-0.01 (0.00) (0.00) (-0.01)	0.00 (0.00) (0.00) (0.00)	0.00 (0.00) (0.00) (-0.02)	-0.01 (-0.04) (-0.01) (0.00)
				DGP SEM			
GBM	0.0	0.3	0.14 (0.18) (0.07) (0.12)	0.39 (0.42) (0.09) (0.42)	0.73 (0.65) (0.04) (0.76)	0.88 (0.84) (0.00) (0.89)	0.92 (0.80) (0.00) (0.89)
dif-Stge			0.07 (0.09) (0.05) (0.04)	0.10 (0.09) (0.04) (0.12)	0.16 (-0.04) (-0.01) (0.14)	0.11 (0.08) (-0.02) (-0.03)	0.05 (-0.07) (0.00) (-0.05)
dif-Gets			0.11 (0.16) (0.07) (0.09)	0.28 (0.22) (0.02) (0.27)	0.34 (0.27) (-0.01) (0.34)	0.30 (0.33) (0.00) (0.16)	0.19 (0.04) (0.00) (0.00)
GBM	0.0	0.5	0.31 (0.25) (0.05) (0.33)	0.82 (0.67) (0.04) (0.84)	0.96 (0.65) (0.00) (0.94)	0.94 (0.58) (0.00) (0.93)	0.96 (0.44) (0.00) (0.97)
dif-Stge			0.07 (0.03) (0.01) (0.05)	0.05 (0.00) (0.02) (0.00)	-0.02 (-0.06) (0.00) (0.00)	-0.03 (-0.07) (0.00) (-0.07)	-0.02 (-0.12) (0.00) (-0.03)
dif-Gets			0.25 (0.21) (0.05) (0.22)	0.38 (0.22) (0.02) (0.29)	0.12 (-0.06) (0.00) (0.07)	-0.01 (-0.20) (0.00) (-0.06)	-0.03 (-0.38) (0.00) (-0.03)
GBM	0.0	0.7	0.66 (0.56) (0.05) (0.60)	0.91 (0.67) (0.04) (0.91)	0.93 (0.31) (0.00) (0.96)	0.95 (0.29) (0.00) (0.97)	0.94 (0.15) (0.00) (0.86)
dif-Stge			0.02 (-0.02) (-0.06) (0.04)	-0.04 (-0.09) (0.00) (-0.03)	-0.04 (-0.22) (0.00) (-0.04)	-0.03 (-0.11) (0.00) (0.00)	-0.03 (-0.05) (0.00) (-0.08)
dif-Gets			0.41 (0.49) (0.05) (0.41)	0.02 (-0.09) (0.00) (0.07)	-0.04 (-0.58) (0.00) (-0.04)	-0.03 (-0.66) (0.00) (0.02)	-0.01 (-0.34) (0.00) (-0.08)
				DGP SDM			
GBM	0.3	0.0	0.05 (0.02) (0.00) (0.05)	0.65 (0.33) (0.05) (0.60)	0.96 (0.95) (0.27) (0.93)	1.00 (0.98) (0.71) (1.00)	1.00 (1.00) (0.90) (1.00)
dif-Stge			0.02 (0.02) (0.00) (0.03)	0.13 (0.04) (0.03) (0.10)	0.05 (0.04) (0.11) (0.02)	0.00 (0.00) (0.09) (0.00)	0.00 (0.00) (0.00) (0.00)
dif-Gets			0.02 (0.02) (0.00) (0.04)	0.14 (0.00) (0.03) (0.12)	0.08 (-0.01) (0.03) (0.02)	0.01 (0.00) (0.07) (0.00)	0.00 (0.00) (-0.04) (0.00)
GBM	0.5	0.0	0.21 (0.15) (0.00) (0.20)	0.92 (0.78) (0.29) (0.90)	1.00 (1.00) (0.65) (1.00)	1.00 (1.00) (0.93) (1.00)	1.00 (0.98) (1.00) (1.00)
dif-Stge			0.07 (0.04) (0.00) (0.08)	0.04 (0.03) (0.07) (0.01)	0.00 (0.00) (0.01) (0.00)	0.00 (0.00) (0.00) (0.00)	0.00 (0.00) (0.00) (0.00)
dif-Gets			0.07 (0.04) (0.00) (0.09)	0.03 (0.07) (0.00) (0.02)	0.00 (0.00) (0.01) (0.00)	0.00 (0.00) (0.00) (0.00)	0.00 (0.00) (0.00) (0.00)
GBM	0.7	0.0	0.27 (0.22) (0.07) (0.36)	0.85 (0.67) (0.35) (0.77)	0.99 (0.95) (0.45) (0.97)	1.00 (1.00) (0.82) (1.00)	1.00 (1.00) (0.96) (1.00)
dif-Stge			0.04 (0.06) (0.02) (0.08)	0.03 (0.05) (0.10) (0.01)	0.01 (0.02) (0.01) (0.00)	0.00 (0.00) (0.06) (0.00)	0.00 (0.00) (0.00) (0.00)
dif-Gets			0.03 (0.02) (0.02) (0.08)	0.03 (0.05) (0.10) (0.01)	0.01 (0.02) (0.01) (0.00)	0.00 (0.00) (0.06) (0.00)	0.00 (0.00) (0.00) (0.00)

The first Figure shows ...dif Stge... shows result of diff(GBM - Stge Strategy) and ...difGets... shows result of diff(GBM - Gets Strategy)

which allow a better approach to reality. At the same time, advances in algorithms and computational techniques make it possible to obtain accurate results in reasonable times. The identification of the ideal specifications requires careful consideration in spatial context and becomes very relevant for obtaining reliable models.

Our simulations highlight the dangers of an incorrect choice of model. Classical probit models ignore spatial dependency effects which produce biased and inconsistent coefficients. The use of SAR and SEM models does not guarantee unbiased parameters, since the presence of spatial structure in the exogenous variables would make us misinterpret the outcomes of the model. Aligned with other researchers (LeSage, 2014b, and Rüttenauer (2019)), it seems that SDM provides the best results. This leads us to the second part of the analysis carried out. One might then think that the best model to start tracking the true underlying generating mechanism behind a binomial variable is the SDP, and therefore follow a General-to-Specific (Gets) strategy. In this research, we compare two novel algorithms applied to spatial probit (General-to-Specific vs Specific-to-General). The reality is that following both decision flows we reach similar and satisfactory accuracy ratio levels under ideal conditions. Only for SEM DGP, the strategy Stge seems to be better. The sample size and the degree of spatial dependence are key in determining the accuracy of the strategies. For sample sizes greater than 400 or spatial dependence intensities greater than 0.4, the precision is in most cases greater than 80%.

The models have been built under a criterion of simplicity and econometric rigor. Currently, in practice, the most used criterion to compare spatial probit models has been the LR. Beron and Vijverberg (2004b) demonstrates the power of the LR in an analysis of MC simulations. We have shown that by combining several tests we can achieve accuracy ratios close to 100%. To guarantee that the chosen combinations are close to an optimum, it is shown that a GBM algorithm with even more tests at its disposal obtains success rates similar to the proposed strategies. Even under non-ideal conditions of endogeneity, GBM shows the same weaknesses as the Stge and Gets strategies. Both strategies seem to resist well the lack of some relevant variable. However, Stge is more sensitive to endogeneity problems when identifying a DGP=SEM and Gets is more sensitive to this same problem when identifying an SLX. In the end, it is hard to make a decision on whether a Gets strategy is preferable to Stge (Florax et al., 2003) or the other way around (Hendry, 2006). It should also be taken into account that Stge by definition requires less computational cost or other aspects such as the analyst's general assessment of the data to take one or the other strategy.

This research aims to continue the debate on model selection in a spatial setting. We open the debate to the spatial probit environment. Although, many questions remain open to new research regarding other types of specifications, the inclusion of new spatial presence detection tests, the setting of other non-ideal conditions or the use of non-standardized adjacency matrices. There is no doubt that the spatial modeling community

is growing and will answer all these questions soon.

4.7 Appendix: Confusion matrix between true DGP and Estimated DGP

Table 4.5: Confusion matrix between true DGP and Estimated DGP for n>=900

		Estimated DGP				
		SIM	SLX	SAR	SEM	SDM
		Stge Strategy				
True DGP	SIM	0.96	0	0.02	0.01	0
	SLX	0	0.97	0	0	0.03
	SAR	0.02	0	0.95	0	0.02
	SEM	0.05	0	0.08	0.85	0.01
	SDM	0	0.01	0	0	0.99
		Gets Strategy				
True DGP	SIM	0.94	0.01	0.02	0.03	0
	SLX	0	0.97	0	0	0.03
	SAR	0	0.02	0.93	0.02	0.03
	SEM	0.05	0	0.02	0.91	0.02
	SDM	0	0.01	0	0	0.99

Concluding Remarks

Processes of a binary nature with spatial dependency require special consideration. Throughout this Thesis we have seen the N-dimensional multivariate normal estimation techniques to solve the spatial Probit problem. It is a complex task to solve, although according to previous research (Calabrese and Elkink, 2014, Novkaniza et al. (2019)) and the results that we show in the first chapter, we can say that the academic solutions provided present a high degree of robustness. The estimation methods detailed in Martinetti and Geniaux (2017) (ML Algorithm) and @ LeSage (2000a) (Gibbs Algorithm) present excellent properties and a very remarkable relationship between precision and computation time. Our simulations indicate that for both ideal and non-ideal conditions, these algorithms obtain centered estimators even when spatial dependence is considerably high and estimation times are acceptable to the analyst.

It is especially important to have algorithms that provide accurate estimates, since it directly impacts the results of the applications using the spatial Probit methodology. As a result, there is an upward trend in publications that adopt this type of technique in many fields. There is no doubt about the growing interest in spatial models. However, the potential of these models still needs to be popularized to give proper answers to problems in human-behavior analytics.

Our applied studies provide interesting discoveries and show the importance of the spatial factor in the permanence of customers in a company. Without the autoregressive component of spatial dependency, we would not achieve efficient results and the coefficients would have a significant bias that can lead to inaccurate and unfocused commercial actions. The results also provide knowledge about the spatial effects on the relationship between the client and the company. Together, they provide valuable information to enable companies to design focused churn reduction.

Another fundamental aspect in the modeling process is the search for the correct specification. It does not have a predominant position in the academic literature. However, a correct selection of the functional form of a model is vital, since the efficiency and unbiasedness of the estimated parameters depend on it. In practice, there is no uniform way to select the true specification of the spatial Probit. It seems that there is no strategy to follow for the selection, but the use of a certain statistic to compare models. The designed strategies Stge and Gets provide a good degree of performance and open the debate on which of them would be the best under non-ideal conditions in the data. In the results of our Monte Carlo analysis there is no clear winner, but depending on the scenario, one is better than the other.

Despite the advances, more research is pending for the future. Advances in computing, parallelization techniques, greater resources in the cloud will make spatial Probit estimation algorithms more accessible. It will be necessary to measure these advances to

shed light on the possibilities and strengthen these techniques. An increase in research with these methods in different areas will also be necessary. Much remains to be said in the field of human behavior. We will also have to put a lot of focus on the selection of spatial Probit models. The sensitivity of the proposed algorithms to changes in W, or the improvement of the strategies in the face of new tests, in all likelihood, will be of great help to research in economics.

Bibliography

Agiakloglou, C. and Tsimpanos, A. (2021). Evaluating information criteria for selecting spatial processes. *The Annals of Regional Science*, 66(3):677–697.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer.

Allahyari, R. and Vahidy, K. (2012). Applying data mining to insurance customer churn management. *IPCSIT*, 30:82–92.

Amaral, A., Abreu, M., and Mendes, V. (2014). The spatial probit model—an application to the study of banking crises at the end of the 1990's. *Physica A: Statistical Mechanics and its Applications*, 415:251–260.

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

Anselin, L. (1988). *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media.

Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical Analysis*, 27(2):93–115.

Anselin, L. (2002). Under the hood issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 27(3):247–267.

Anselin, L. (2003). Geoda 0.9 user's guide. *Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL*.

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science*, 89(1):3–25.

Anselin, L., Bera, A. K., Florax, R., and Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, 26(1):77–104.

Anselin, L. and Florax, R. (2012). *New directions in spatial econometrics*. Springer Science & Business Media.

Anselin, L. and Li, X. (2019). Operational local join count statistics for cluster detection. *Journal of Geographical Systems*, 21(2):189–210.

Antonini, G., Bierlaire, M., and Weber, M. (2006). Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8):667–687.

Arbia, G. (2014). A primer for spatial econometrics with applications in r.

Archaux, C., Martin, A., and Khenchaf, A. (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 and Communication Technologies.

Arima, E. (2016). A spatial probit econometric model of land change: the case of infrastructure development in western amazonia, peru. *PloS one*, 11(3):e0152058.

Autant-Bernard, C. (2006). Where do firms choose to locate their r & d? *A spatial conditional logit analysis on French data, European Planning Studies*, 14:1187–1208.

Autant-Bernard, C., Lesage, J. P., and Parent, O. (2007). Firm innovation strategies: a spatial cohort multinomial probit approach. *Annales d'Economie et de Statistique*, pages 63–80.

Avery, R. B., Hansen, L. P., and Hotz, V. J. (1983). Multiperiod probit models and orthogonality condition estimation. *International Economic Review*, pages 21–35.

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

Baidoo, I. K. and Nyarko, E. (2015). A discrete choice modeling of service quality attributes in public transport. *Research Journal of Mathematics and Statistics*, 7(1):6–10.

Basile, R., Durbán, M., Mínguez, R., Montero, J. M., and Mur, J. (2014). Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control*, 48:229–245.

Beron, K. J. and Vijverberg, W. P. (2004a). Probit in a spatial context: a monte carlo analysis. In *Advances in Spatial Econometrics*, pages 169–195. Springer.

Beron, K. J. and Vijverberg, W. P. (2004b). Probit in a spatial context: a monte carlo analysis. In *Advances in Spatial Econometrics*, pages 169–195. Springer.

Billé, A. G. and Arbia, G. (2019). Spatial limited dependent variable models: A review focused on specification, estimation, and health economics applications. *Journal of Economic Surveys*, 33(5):1531–1554.

Bivand, R., Millo, G., and Piras, G. (2021). A review of software for spatial econometrics in r. *Mathematics*, 9(11):1276.

Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., and Pebesma, E. J. (2008). *Applied spatial data analysis with R*, volume 747248717. Springer.

Bliss, C. I. (1934). The method of probits. *Science*, 79(2037):38–39.

Bliss, C. I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22(1):134–167.

Box, G. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society*, 143:383–430.

Brasington, D., Flores-Lagunes, A., and Guci, L. (2016). A spatial model of school district open enrollment choice. *Regional Science and Urban Economics*, 56:1–18.

Brockett, P. L., Golden, L. L., Guillen, M., Nielsen, J. P., Parner, J., and Perez-Marin, A. M. (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., Cook, A. R., Lee, J. S. H., Wich, S. A., Koh, L. P., and Carrasco, L. R. (2015). Analysis of deforestation and protected area effectiveness in indonesia: a comparison of Bayesian spatial models. *Global Environmental Change*, 31:285–295.

Buckinx, W. and Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting. *European Journal of Operational Research*, 164(1):252–268.

Burez, J., , and den Poel, V. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3):4626–4636.

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

Calabrese, R. and Elkink, J. A. (2014). Estimators of binary spatial autoregressive models: A monte carlo study. *Journal of Regional Science*, 54(4):664–687.

Case, A. C. (1992). Neighborhood influence and technological change. *Regional Science and Urban Economics*, 22:491–508.

Cattaneo, L. and Rizzolatti, G. (2009). The mirror neuron system. *Archives of neurology*, 66(5):557–560.

Chartrand, T. L. and Bargh, J. A. (1999). The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893.

Chasco, C. (2003). *Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales*. Dirección General de Economía y Planificación.

Chu, B. H., Tsai, M. S., and Ho, C. S. (2007). Toward a hybrid data mining model for customer retention. *Knowledge-Based Systems*, 20(8):703–718.

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

Cliff, A. and Ord, J. (1973). Spatial autocorrelation.

Cliff, A. and Ord, J. K. (1981). *Spatial processes Models and applications*. Pion.

Cliff, A. D. (1973). Spatial autocorrelation.

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

Coussement, K. and den Poel, D. V. (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.

Cox, D. R. (1969). *Analysis of binary data*. Chapman and Hall.

Cramer, J. S. (2002). The origins of logistic regression.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Davidson, R., MacKinnon, J. G., et al. (2004). *Econometric theory and methods*, volume 5. Oxford University Press New York.

De la Llave, M. and López, F. (2020). Spatial models for online retail churn: Evidence from an online grocery delivery service in madrid. *Papers in Regional Science*, 99(6):1643–1665.

De la Llave, M., López, F., and Angulo, A. (2019a). The impact of geographical factors on churn prediction: an application to an insurance company in madrid's urban area. *Scandinavian Actuarial Journal*, 3:188–203.

De la Llave, M. Á., López, F. A., and Angulo, A. (2019b). The impact of geographical factors on churn prediction: an application to an insurance company in madrid's urban area. *Scandinavian Actuarial Journal*, 2019(3):188–203.

de la Llave Montiel, M. A. and López, F. (2020). Spatial models for online retail churn: Evidence from an online grocery delivery service in madrid. *Papers in Regional Science*, 99(6):1643–1665.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977a). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977b). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Dierkes, T., Bichler, M., and Krishnan, R. (2011). Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with Markov logic networks. *Decision Support Systems*, 51(3):361–371.

Droftina, U., Štular, M., and Košir, A. (2015). A diffusion model for churn prediction based on sociometric theory. *Advances in Data Analysis and Classification*, 9(3):341–365.

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

Elhorst, J. P. et al. (2014). *Spatial econometrics: from cross-sectional data to spatial panels*, volume 479. Springer.

Elhorst, J. P. and Halleck Vega, S. (2017). The slx model: extensions and the sensitivity of spatial spillovers to w. *Papeles de Economía Española*, 152:34–50.

Elhorst, P., Heijnen, P., Samarina, A., and Jacobs, J. (2013). State transfers at different moments in time: A spatial probit approach. *University of Groningen, Faculty of Economics and Business*.

Elms, J., De Kervenoael, R., and Hallsworth, A. (2016). Internet or store? an ethnographic study of consumers' internet and store-based grocery shopping practices. *Journal of Retailing and Consumer Services*, 32:234–243.

Ferreira, P., Telang, R., and De Matos, M. G. (2019). Effect of friends' churn on consumer behavior in mobile networks. *Journal of Management Information Systems*, 36(2):355–390.

Fiorio, C. V., Florio, M., and Perucca, G. (2013). User satisfaction and the organization of local public transport: Evidence from european cities. *Transport Policy*, 29:209–218.

Florax, R. and Folmer, H. (1992). Specification and estimation of spatial linear regression models: Monte carlo evaluation of pre-test estimators. *Regional Science and Urban Economics*, 22(3):405–432.

Florax, R., Folmer, H., and Rey, S. (2003). Specification searches in spatial econometrics: the relevance of hendry's methodology. *Regional Science and Urban Economics*, 33(5):557–579.

Fortin, M., Delisle-Boulianne, S., and Pothier, D. (2013). Considering spatial correlations between binary response variables in forestry: an example applied to tree harvest modelling. *Forest Science*, 59:253–260.

Franses, P. H. and Paap, R. (2001). *Quantitative Models in Marketing Research (Cambridge*. Cambridge University Press, UK.

Frasquet Deltoro, M., Molla Descals, A., and Ruiz Molina, M. E. (2012). Factores determinantes y consecuencias de la adopción del comercio electrónico b2c: una comparativa internacional. *Estudios Gerenciales*, 28(123):101–120.

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

Gaddum, J. (1933). Reports on biological standards. iii. methods of biological assay depending on a quantal response. spec. report series, med. res. council, 183, his. *Maj. Sta. Off., London*.

Gallego, M. D., Bueno, S., and Terreño, J. F. (2016). Motivations and barriers to set up e-commerce in spain: A delphi study. *Estudios Gerenciales*, 32(140):221–227.

Giansoldati, M., Rotaris, L., Scorrano, M., and Danielis, R. (2020). Does electric car knowledge influence car choice? evidence from a hybrid choice model. *Research in Transportation Economics*, 80:100826.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*.

Goetzke, F. and Andrade, P. M. (2010). Walkability as a summary measure in a spatially autoregressive mode choice model: an instrumental variable approach. In *Progress in Spatial Analysis*, pages 217–229. Springer.

Gómez-Rubio, V., Bivand, R. S., and Rue, H. (2021). Estimating spatial econometrics models with integrated nested laplace approximation. *Mathematics*, 9(17):2044.

Greene, W. (2012). Econometric analysis. 7th (international) ed. new york university. *Pearson. ISBN*, 13:978–0.

Greenwell, B., Boehmke, B., Cunningham, J., and Developers, G. (2020). *gbm: Generalized Boosted Regression Models*. R package version 2.1.8.

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

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

Haghani, M., Bliemer, M. C., and Hensher, D. A. (2021). The landscape of econometric discrete choice modelling research. *Journal of Choice Modelling*, 40:100303.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–318.

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

Hendry, D. (1979). Predictive failure and econometric modelling in macroeconomics: the transactions demand for money. In Ormerod, P., editor, *Economic Modelling*. Heinemann, London.

Hendry, D. F. (2006). A comment on "specification searches in spatial econometrics: The relevance of hendry's methodology". *Regional Science and Urban Economics*, 36(2):309–312.

Herrera-Gomez, M., Matilla-Garcia, M., and Ruiz-Marin, M. (2021). Spatial partial causality. *The European Physical Journal Special Topics*, pages 1–5.

Holloway, G., Shankar, B., and Rahmanb, S. (2002). Bayesian spatial probit estimation: a primer and an application to hyv rice adoption. *Agricultural Economics*, 27(3):383–402.

Hsu, M. K., Huang, Y., and Swanson, S. (2010). Grocery store image, travel distance, satisfaction and behavioral intentions. *International Journal of Retail & Distribution Management*, 38(2):115–132.

Huigevoort, C. and Dijkman, R. (2015). *Customer churn prediction for an insurance company*. PhD thesis, Thesis.

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

Iacoboni, M. (2005). Neural mechanisms of imitation. *Current opinion in neurobiology*, 15(6):632–637.

İlhan, B. Y. and İşçioğlu, T. E. (2015). Effect of women's labor market status on online grocery shopping, the case of turkey. *Eurasian Business Review*, 5(2):371–396.

Kahle, D. and Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161.

Kee, H. T. and Wan, D. (2004). Intended usage of online supermarkets: The singapore case. In *In The Fourth International Conference on Electronic Business*, pages 1308–1312.

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

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

Kelejian, H. H. and Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1):53–67.

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

Klier, T. and McMillen, D. P. (2008). Clustering of auto supplier plants in the united states: generalized method of moments spatial logit for large samples. *Journal of Business & Economic Statistics*, 26(4):460–471.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496.

Lacombe, D. and LeSage, J. P. (2015). Use and interpretation of spatial autoregressive probit models. *Annals Regional Science*, 60(1):1–24.

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

Lai, Y. and Zeng, J. (2014). Analysis of customer churn behavior in digital libraries. *Program*.

Läpple, D., Holloway, G., Lacombe, D. J., and O'Donoghue, C. (2017). Sustainable technology adoption: a spatial analysis of the irish dairy sector. *European Review of Agricultural Economics*, 44(5):810–835.

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

Leamer, E. E. and Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*, volume 53. John Wiley & Sons Incorporated.

Leenders, R. (2002). Modeling social influence through network autocorrelation: constructing the weight matrix. *Social Networks*, 24(1):21–47.

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

LeSage, J. and Pace, R. K. (2009a). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

LeSage, J. and Pace, R. K. (2009b). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

LeSage, J. P. (2000a). Bayesian estimation of limited dependent variable spatial autoregressive models. *Geographical Analysis*, 32(1):19–35.

LeSage, J. P. (2000b). Bayesian estimation of limited dependent variable spatial autoregressive models. *Geographical Analysis*, 32(1):19–35.

LeSage, J. P. (2014a). What regional scientists need to know about spatial econometrics. *Available at SSRN 2420725*.

LeSage, J. P. (2014b). What regional scientists need to know about spatial econometrics. *Available at SSRN 2420725*.

LeSage, J. P., Kelley Pace, R., Lam, N., Campanella, R., and Liu, X. (2011). New orleans business recovery in the aftermath of hurricane katrina. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(4):1007–1027.

Lochl, M., Hauri, H. R., and Axhausen, K. W. (2009). Agents, space and market shares: a spatial analysis of the swiss insurance market. *Arbeitsberichte Verkehrs-und Raumplanung*, 557.

Malthus, T. (1798). An essay on the principle of population. an essay on the principle of population, as it affects the future improvement of society with remarks on the speculations of mr. godwin, m. condorcet, and other writers. *St. Paul's church-yard*, page 4.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542.

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

Martinetti, D. and Geniaux, G. (2021). *ProbitSpatial: Probit with Spatial Dependence, SAR, SEM and SARAR Models*. R package version 1.1.

Martínez, M. M. and Vázquez, M. S. (2008). Decisive parameters of the online purchase in the virtual supermarkets. pages 475–478.

Mate-Sánchez-Val, M. (2021). The impact of geographical positioning on agri-food businesses' failure considering nonlinearities. *Agribusiness*, 37(3):612–628.

McFadden, D. et al. (1973). Conditional logit analysis of qualitative choice behavior.

McMillen, D. (2013). *McSpatial: Nonparametric spatial data analysis*. R package version 2.0.

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

McMillen, D. P. and McDonald, J. F. (2002). Land values in a newly zoned city. *Review of Economics and Statistics*, 84(1):62–72.

Melnyk, V., Van Osselaer, S. M., and Bijmolt, T. H. (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.

Meltzoff, A. N. and Decety, J. (2003). What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431):491–500.

Mendell, N. and Elston, R. (1974a). Multifactorial qualitative traits: genetic analysis and prediction of recurrence risks. *Biometrics*, 30:41–57.

Mendell, N. R. and Elston, R. (1974b). Multifactorial qualitative traits: genetic analysis and prediction of recurrence risks. *Biometrics*, pages 41–57.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Miguéis, V. L., Van den Poel, D., Camanho, A. S., and e Cunha, J. F. (2012). Predicting partial customer churn using markov for discrimination for modeling first purchase sequences. *Advances in Data Analysis and Classification*, 6(4):337–353.

Milborrow, S. (2011). Derived from mda: mars by t. hastie and r. tibshirani. *Earth: Multivariate Adaptive Regression Splines*, 2011.

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

Moffitt, R. A. (2001). Policy interventions, low-level equilibria, and social interactions. *Social Dynamics*, 4(45-82):6–17.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

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

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., and Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, 11(3):690–696.

Mur, J. and Angulo, A. (2009). Model selection strategies in a spatial setting: Some additional results. *Regional Science and Urban Economics*, 39(2):200–213.

Naus, J. (1974). Probabilities for a generalized birthday problem. *Journal of the American Statistical Association*, 69(347):810–815.

Naus, J. L. (1965). Clustering of random points in two dimensions. *Biometrika*, 52(1-2):263–266.

Nazari Gooran, A. and Borimnejad, V. (2015). Bayesian analysis of spatial probit models for investigating the adoption of high yielding wheat varieties. *Economic Modeling*, 7(21):69–83.

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

Nilsson, E., G"arling, T., Marell, A., and Nordvall, A. C. (2015). Who shops groceries where and how?–the relationship between choice of store format and type of grocery shopping. the international review of retail. *Distribution and Consumer Research*, 25(1):1–19.

Novkaniza, F., Djuraidah, A., Fitrianto, A., and Sumertajaya, I. (2019). Simulation study for comparison of spatial autoregressive probit estimation methods. In *IOP Conference Series: Earth and Environmental Science*, volume 299, page 012030. IOP Publishing.

Ommani, A. and Noorollah Noorivandi, A. (2019). Bayesian analysis of spatial probit models in wheat waste management adoption. *International Journal of Agricultural Management and Development*, 9(1):37–44.

Ortega-García, J. A., López-Hernández, F. A., Cárceles-Álvarez, A., Fuster-Soler, J. L., Sotomayor, D. I., and Ramis, R. (2017). Childhood cancer in small geographical areas and proximity to air-polluting industries. *Environmental research*, 156:63–73.

Ortega-García, J. A., López-Hernández, F. A., Funes, M. L. A., Sauco, M., and Ramis, R. (2020). My partner and my neighbourhood: The built environment and social networks' impact on alcohol consumption during early pregnancy. *Health and Place*, 61(10223):9.

Pace, R. K. and LeSage, J. P. (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: 0*, 1(2016):3–34.

Paelinck, j. and Klaassen, L. (1979). *Spatial autocorrelation*. Farnborough: Saxon House.

Páez, A., López, F. A., Ruiz, M., and Morency, C. (2013). Development of an indicator to assess the spatial fit of discrete choice models. *Transportation Research Part B: Methodological*, 56:217–233.

Park, C. H. (2017). Online purchase paths and conversion dynamics across multiple websites. *Journal of Retailing*, 93(3):253–265.

Pearl, R. and Reed, L. J. (1920). On the rate of growth of the population of the united states since 1790 and its mathematical representation. *Proceedings of the National Academy of Sciences of the United States of America*, 6(6):275.

Pinheiro, C. and Helfert, M. (2010). Neural network and social network to enhance the customer loyalty process. In *Innovations and Advances in Computer Sciences and Engineering (Netherlands*, pages 91–96. Springer.

Pinkse, J. (1999). Asymptotics of the moran test and a test for spatial correlation in probit models. Technical report, Department of Economics Working Paper.

Pinkse, J. (2004). Moran-flavored tests with nuisance parameters: examples. In *Advances in Spatial Econometrics*, pages 67–77. Springer.

Pinkse, J. and Slade, M. E. (1998). Contracting in space: An application of spatial statistics to discrete-choice models. *Journal of Econometrics*, 85(1):125–154.

Reichheld, F. F. and Sasser, W. E. (1990). Zero defections: Quality comes to services. *Harvard Business Review*, 68(5):105–111.

Richards, T. J., Hamilton, S. F., and Allender, W. J. (2014). Social networks and new product choice. *American Journal of Agricultural Economics*, 96(2):489–516.

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

Rizzolatti, G. and Craighero, L. (2004). The mirror-neuron system. *Annu. Rev. Neurosci.*, 27:169–192.

Rodríguez, C., Sánchez-Val, M. M., and Hernández, F. A. L. (2016). The geographic proximity in the spillover effects of business failure in smes: Empirical application with the spatial probit model. *Studies of Applied Economics*, 34(3):619–638.

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

Ruiz, M., López, F., and Páez, A. (2010). Testing for spatial association of qualitative data using symbolic dynamics. *Journal of Geographical Systems*, 12(3):281–309.

Rüttenauer, T. (2019). Spatial regression models: a systematic comparison of different model specifications using monte carlo experiments. *Sociological Methods & Research*, page 0049124119882467.

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

Schwarz, G. (1978). Estimating the dimension of a model the annals of statistics 6 (2), 461–464. *URL: http://dx. doi. org/10.1214/aos/1176344136*.

Scott, L. M. and Janikas, M. V. (2010). Spatial statistics in arcgis. In *Handbook of applied spatial analysis*, pages 27–41. Springer.

Scott Long, J. (1997). Regression models for categorical and limited dependent variables. *Advanced quantitative techniques in the social sciences*, 7.

Skevas, T., Skevas, I., and Kalaitzandonakes, N. (2021). The role of peer effects on farmers' decision to adopt unmanned aerial vehicles: evidence from missouri. *Applied Economics*, pages 1–11.

Soeini, R. and Rodpysh, K. (2012). Applying data mining to insurance customer churn management. *International Proceedings of Computer Science*, 30:82–92.

Srinivasan, S. S., Anderson, R., and Ponnavolu, K. (2002). Customer loyalty in e-commerce: an exploration of its antecedents and consequences. *Journal of Retailing*, 78(1):41–50.

StataCorp, L. (2017). Stata spatial autoregressive models reference manual.

Storm, H., Mittenzwei, K., and Heckelei, T. (2015). Direct payments, spatial competition, and farm survival in norway. *American Journal of Agricultural Economics*, 97(4):1192–1205.

Theil, H. (1969). A multinomial extension of the linear logit model. *International Economic Review*, 10(3):251–259.

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

Trivedi, M. (2011). Regional and categorical patterns in consumer behavior: revealing trends. *Journal of Retailing*, 87(1):18–30.

Tsai, C. F. and Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, 36(10):12547–12553.

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

Valle, D., Hyde, J., Marsik, M., and Perz, S. (2020). Improved inference and prediction for imbalanced binary big data using case-control sampling: A case study on deforestation in the amazon region. *Remote Sensing*, 12(8):1268.

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

Verhulst, P.-F. (1838). Notice sur la loi que la population suit dans son accroissement. *Corresp. Math. Phys.*, 10:113–126.

Wang, C.-H., Akar, G., and Guldmann, J.-M. (2015). Do your neighbors affect your bicycling choice? a spatial probit model for bicycling to the ohio state university. *Journal of Transport Geography*, 42:122–130.

Wang, H., Iglesias, E. M., and Wooldridge, J. M. (2013). Partial maximum likelihood estimation of spatial probit models. *Journal of Econometrics*, 172(1):77–89.

Wang, X. and Kockelman, K. M. (2009). Application of the dynamic spatial ordered probit model: Patterns of land development change in austin, texas. *Papers in Regional Science*, 88(2):345–365.

Wang, Y., Kockelman, K. M., and Wang, X. (2011). Anticipation of land use change through use of geographically weighted regression models for discrete response. *Journal of the Transportation Research Board*, 2245:111–123.

Wilhelm, S. and de Matos, M. G. (2013). Estimating spatial probit models in r. *R J.*, 5(1):130.

Wilhelm, S. and de Matos, M. G. (2015). *spatialprobit: Spatial Probit Models*. R package version 0.9-11.

Wood, A. (2011). *Multichannel shopping: Crossing the channels*. AdMap, Warc.

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

Yang, S. and Allenby, G. M. (2003a). Modeling interdependent consumer preferences. *Journal of Marketing Research*, 40(3):282–294.

Yang, S. and Allenby, G. M. (2003b). Modeling interdependent consumer preferences. *Journal of Marketing Research*, 40(3):282–294.

Yang, W. and Knook, J. (2021). Spatial evaluation of the impact of a climate change participatory extension programme on the uptake of soil management practices. *Australian Journal of Agricultural and Resource Economics*, 65(3):539–565.

Yang, W. and Sharp, B. (2017). Spatial dependence and determinants of dairy farmers' adoption of best management practices for water protection in new zealand. *Environmental Management*, 59(4):594–603.

Yule, G. (1925). *The growth of population and the factors which control it*. Harrison & Sons.

Zhang, X., Zhu, J., Xu, S., and Wan, Y. (2012). Predicting customer churn through interpersonal influence. *Knowledge-Based Systems*, 28:97–104.

Zheng, H., Ma, W., and Li, G. (2021). Learning from neighboring farmers: Does spatial dependence affect adoption of drought-tolerant wheat varieties in china? *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*.