Mortgage default decisions in the presence of non-normal, spatially dependent disturbances

Mortgage Default Decisions in the Presence of Non-normal, Spatially Dependent Disturbances

Raffaella Calabrese[1]
Business School, University of Edinburgh
raffaella.calabrese@ed.ac.uk
and
Meagan McCollum
Zicklin School of Business, Baruch College
meagan.mccollum@baruch.cuny.edu
and
R. Kelley Pace
LREC Endowed Chair of Real Estate
Department of Finance, Louisiana State University
kelley@spatial.us

[1]The author would like to acknowledge support for this research provided by Small Grant of the British Academy and the Regional Studies Association Membership Research Grant.

Abstract

We develop a flexible binary choice model for mortgage default decisions that incorporates neighborhood effects in the disturbances. The main advantage of the model lies in its performance in providing accurate estimates of the probability of default for risky mortgage loans. In addition, it can be applied to portfolios with a high number of loans. Assuming mortgage decisions with spatially dependent disturbances, the proposed approach uses the generalized extreme value distribution to flexibly model the error terms. To estimate the model on a large sample size, we use a variant of the Geweke-Hajivassiliou-Keane algorithm. We apply the proposed model and its competitors to a large dataset on almost 300,000 mortgages in Clark County, which includes Las Vegas, over 2009-2010. The results show that our proposal greatly improves the predictive accuracy of identifying loans that will default. Moreover, the competitor models underestimate credit Value at Risk.

Keywords: binary imbalanced samples, spatial econometrics, generalized extreme value distribution, mortgage default decisions.

1 Introduction

Problems emanating from the mortgage market played a role in the Great Recession and has demonstrated the importance of better modeling of household mortgage default. The literature on mortgage default has emphasized the role of house prices as well as home equity accumulation for the default decision (Deng et al. 2000; Ghent and Kudlyak, 2011; Mayer et al. 2009; Mian et al. 2010; Scharlemann and Shore, 2016; Zhu and Pace, 2015). Recently, Scharlemann and Shore (2016) have examined the effect of negative equity on borrowers' mortgage default under the Principal Reduction Alternative (PRA), part of the government's Home Affordable Modification Program (HAMP), which was introduced to reduce mortgage payments of borrowers with negative equity who are likely to default.

Although existing studies have established the importance of modeling mortgage risk, the risk associated with neighborhood effects in the disturbances is under-explored. Agarwal et al. (2012) have examined how the concentration in the same zip code of defaulted mortgage affected individual loan performance, finding some significant neighborhood effects. The authors account for neighborhood effects by including zip code fixed effects corresponding to property location. Harding et al. (2009) have shown that foreclosures reduce the prices of nearby non-distressed sales through a neighborhood effect. These effects could arise because of the neglect of vacant properties or as a consequence of the reduction in maintenance of properties by defaulted borrowers.

Although fixed effects provide the most common way to model such neigh-

borhood effects, the very local nature of real estate requires a large number of fixed effects. As an alternative, recently spatial autoregressive models been used to model neighborhood effects in mortgage defaults. Zhu and Pace (2014) have investigated spatial dependence in the disturbances and the effect of borrower characteristics from nearby properties on own default propensity. They find that allowing spatial dependence in the disturbances greatly improves the predictive accuracy of credit risk models.

This result is a consequence of the influence of neighbors' characteristics on a borrower's propensity to default on a mortgage. For example, if the houses in a neighborhood are in a poor condition the expectation of future appreciation is low, leaving the borrower with less of an incentive to repay her mortgage.

Lenders employ credit standards that ordinarily result in low levels of default. In other words, most times a default is a rare event and the estimated probability of a default depends partially on the assumed error distribution. Using a spatial probit model, Zhu and Pace (2014) assume that the errors are normally distributed, which gives little weight to rare events. As the distribution of the errors in this model is symmetric, borrowers are subject to approximately equal levels of positive and negative random influences in their decisions. However, the omission of relevant skewed variables, such as wealth, could lead to an overall error composed of both a symmetric component and a skewed component. Therefore, the overall error could be skewed.

In addition, logit models often rely on a utility justification where each choice has associated with a Gumbel distributed error (McFadden, 1978). The utility difference between two choices has a systematic part and an error part,

composed of the difference between two Gumbel distributed errors, which leads to the symmetric logit distribution. However, if the variance of the errors associated with each choice differs, this could also lead to a skewed distribution of the overall error.[1] If we model mortgage decisions, disturbances can be spatially dependent because location related variables can be omitted and because nearby properties show similar values for those omitted variables.

Furthermore, one could argue for using a distribution with some tail weight on decision theoretic grounds. If a rare event, such as a default, has associated with it a much larger cost (loss) than the benefit (profit) associated with the common event of loan repayment it would argue for using a method which has substantial tail weight.

In a non-spatial context, various papers (Calabrese et al. 2015; King and Zeng, 2001; and Wang and Dey, 2010) have dealt with this rare binary event problem. Different methods have been proposed to overcome the challenges associated with this problem. Over-sampling rare events and under-sampling common events have been proposed (for a review see Sahare and Gupta, 2012), but this approach encounters difficulty when applied to spatial data since it alters the spatial structure of the data and can potentially change the estimates of spatial spillovers and spatial dependence. Another approach suggested in the literature, although not in a spatial context, is to use a flexible skewed link function such as one based on the Generalized Extreme Value (GEV) distribution. This approach effectively increases the weight given to the rare event (Calabrese et al. 2015; Wang and Dey, 2010).

[1]Such heteroscedastic choice models across alternatives have often been employed in political science where interest lies in estimating the uncertainty associated with the choices (Zeng, 2000). As an example, consider the choice between Donald Trump and Hilary Clinton.

Calabrese and Elkink (2016) applied the GEV approach to spatial data, but encountered computational problems when using Gibbs sampling for large sample sizes. Insofar as many practical problems involving loan data have a large number of observations, this is a limitation. In contrast, Pace and LeSage (2016) proposed a method to handle binary spatial problems for large sample sizes using the Geweke-Hajivassiliou-Keane algorithm, but they used normal errors which give low weights to rare events.

The contribution of this article is twofold. From a methodological perspective, we propose a spatial choice model suitable for highly imbalanced binary large sample size data. The distribution of the error terms is allowed to be asymmetric and its tail behavior is flexibly determined from the data. Particularly, we assume that the joint distribution of the error terms is a multivariate Generalized Extreme Value (GEV) random variable, whose marginal distributions are also GEV.[2] The advantage of the GEV model we discuss here is that it incorporates a wide range of skewness and kurtosis with the unconstrained shape parameter τ.

From an empirical point of view, we improve the classification performance obtained using classical alternatives for mortgage scoring assessments. We analyze a dataset of almost 300,000 mortgages over 2009-2010 in Clark County, which includes Las Vegas, the city with the largest concentration of subprime mortgages in the US. We show that ignoring neighborhood spillover effects and using logit or probit choice models yield misleading results. For example,

[2]The model we introduce here is totally different from the generalized extreme value models initiated by McFadden (1974). In McFadden's definition, the GEV distribution is Type I extreme value distribution or Gumbel distribution (McFadden, 1978), which is a special case of the GEV distribution we use in the equation (17) when the shape parameter $\tau \to 0$. Furthermore, our proposal generalizes McFadden's model to allow heteroscedasticity across choice alternatives.

in 2009, at the beginning of the foreclosure crisis in Las Vegas, a logit choice model under the assumption of independent mortgage decisions leads to the minimum estimated probability of repayment of 0.9. If we introduce spatially dependent disturbances in a logit model the minimum estimate is 0.73. The Fast Binary Spatial Generalized Extreme Value (FBSGEV) model proposed here achieves a more realistic minimum estimated probability of repayment equal to 0.25.

The paper is organized as follows. The next section reviews the widely used specifications of the binary choice models with spatial dependence. In Section 3 we propose the FBSGEV choice model. Section 4 shows the results obtained from applying the traditional and proposed approaches to data on mortgage decisions. The last section reviews the key findings.

2 Binary Choice Models

We have a portfolio of n mortgages. A borrower labeled i is a decision maker facing two mutually exclusive and collectively exhaustive alternatives – payment of mortgage debt p (indicated by $Y_i = 0$) or default d ($Y_i = 1$). A binary choice model specifies the probability of choosing each alternative as a function of observable variables and unknown parameters to be estimated from sample data. The estimated model can then be used to explain and predict choice behavior.

In the utility maximization approach, the i decision-maker chooses the alternative $j = d,p$ that provides the greatest utility U_{ij}. The dependent variable Y_i can be represented as a latent response model

$$Y_i = \begin{cases} 1, & U_{id} - U_{ip} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The utility that the borrower obtains from the j-th alternative is decomposed into a part that is known up to some parameters $\mathbf{x}_i\boldsymbol{\beta}_j$ and an unknown part ϵ_{ij} that is treated as random

$$U_{ij} = \mathbf{x}_i\boldsymbol{\beta}_j + \epsilon_{ij} \quad \text{for } j = p,d. \tag{2}$$

The borrower's default probability is given by the probability that the decision-maker i chooses the alternative d

$$\begin{aligned} P_{id} &= Prob\{Y_i = 1\} = Prob\{U_{id} > U_{ip}\} = Prob\{\mathbf{x}_i\boldsymbol{\beta}_d + \epsilon_{id} > \mathbf{x}_i\boldsymbol{\beta}_p + \epsilon_{ip}\} \\ &= \int_{\epsilon_p}\int_{\epsilon_d} \mathbf{1}\{\epsilon_{id} - \epsilon_{ip} > \mathbf{x}_i(\boldsymbol{\beta}_d - \boldsymbol{\beta}_p)\} f(\epsilon_{ip}) f(\epsilon_{id}) d\epsilon_{ip} d\epsilon_{id} \end{aligned} \tag{3}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, equaling 1 when the term in parentheses is true and 0 otherwise. The bidimensional integral in the equation (3) is computed over the density of the unobserved portion of utility $f(\cdot)$ under the assumption that ϵ_{ij} are identically and independently distributed. Different discrete choice models are obtained from different specifications of the error terms ϵ_{ij}.

The most widely used binary choice model is logit. Originally, Luce (1959) derived the logit equation from some assumptions about the characteristics of choice probabilities, known as the independence from irrelevant alternatives. McFadden (1974, 1978) extended this analysis assuming that ϵ_{ij} are distributed

as a type I extreme value (Gumbel) and deriving the logit closed-form under this assumption. As the unobserved component ϵ_{ij} has variance $\sigma^2(\pi^2/6)$ in McFadden's model, therefore the logit choice model implies homoscedasticity across choice alternatives.

2.1 Spatial Binary Choice Models

In house price models, disturbances often display statistically significant spatial dependence (e.g. LeSage and Pace, 2004). The mortgage literature, despite relying on house prices, usually assume independent error components, ignoring neighborhood effects. However, Zhu and Pace (2014) have found that the predictive accuracy of a default model is greatly improved when allowing spatial dependence in the disturbances. Thus, we add spatial random effects in the binary choice model to account for latent and unmeasured effects that are spatially structured.

Let $\boldsymbol{\epsilon}_j$ be the n-dimensional vector of disturbances for the alternative j. Spatial interdependence can be introduced in the error terms $\boldsymbol{\epsilon}_j$ as follows

$$\boldsymbol{\epsilon}_j = A\mathbf{v}_j, \tag{4}$$

where $\mathbf{v}_j$ is a vector of independent and identically distributed error terms. Different specifications for the matrix A have been used in the literature (LeSage and Pace, 2009)

- if $A = (I - \rho W)^{-1}$, the model is known as spatial error model (SEM);

- if

$$A = (I - \rho W)^{-1/2}, \tag{5}$$

  the model is known as conditional autoregressive model (CAR);

- if $A = (I + \rho W)$, the model is known as a moving average model (MA),

where W is an exogenous square matrix W of order n and ρ is the associated scalar parameter. The generic element w_{ij} is equal to a positive number when observation j is a neighbor to observation i and 0 otherwise. Neighborhood can refer to geographical or alternative vicinity. In practice, W is often scaled to have a maximum eigenvalue of 1, which simplifies the setting of the interval for the spatial dependence parameter. For a symmetric W which has real eigenvalues, one can either divide a candidate weight matrix by its maximum eigenvalues so that the new matrix has an eigenvalue of 1 or scale the weight matrix so that both the rows and columns sum to 1.[3]

Substituting the equation (4) in the utility function (2), we obtain

$$\boldsymbol{U}_j = X\boldsymbol{\beta}_j + A\mathbf{v}_j \quad \text{for } j = p,d. \tag{6}$$

$$\boldsymbol{U}_d - \boldsymbol{U}_p = X(\boldsymbol{\beta}_d - \boldsymbol{\beta}_p) + A\mathbf{v}_d - A\mathbf{v}_p. \tag{7}$$

Different methods have been proposed to estimate the parameters in the equation (6). Some of the widely used approach are the Gibbs Sampling (LeSage, 2000), the Recursive Importance Sampling (Beron and Vijverberg, 2004) and the Generalized Method of Moments (Pinkse and Slade, 1998; Klier

[3]This matrix becomes doubly stochastic (in the linear algebra sense), although all the entries are non-stochastic (in the statistical sense).

and McMillen, 2008). For a review and comparison of these methods, see Calabrese and Elkink (2014), LeSage and Pace (2009).

The possibility of different levels of spatial dependence between choices is a specification issue with spatial discrete choice models that does not arise in non-spatial models. We examine the utility differences between choices d and p as captured by the $n \times 1$ vector u where, for simplicity, the individual utilities follow a moving average process with different levels of spatial dependence ρ_d, ρ_p as in (8). In (9) we assume *iid* choice utility variances and no covariances among the individual choice utilities. The resulting variance-covariance matrix Ω in (10) through (12) shows that the utility differences still follow a moving average process, but with a different level of dependence ρ_a and variance σ_a^2 as shown in (13).

$$u = \boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p = (I_n + \rho_d W)^{1/2}\mathbf{v}_d - (I_n + \rho_p W)^{1/2}\mathbf{v}_p \tag{8}$$

$$E(\mathbf{v}_d\mathbf{v}_p') = 0_n,\ E(\mathbf{v}_d\mathbf{v}_d') = \sigma_d^2 I_n,\ E(\mathbf{v}_p\mathbf{v}_p') = \sigma_p^2 I_n \tag{9}$$

$$\Omega = E(uu') = (I_n + \rho_d W)\sigma_d^2 + (I_n + \rho_p W)\sigma_p^2 \tag{10}$$

$$= I_n(\sigma_d^2 + \sigma_p^2) + (\rho_d\sigma_d^2 + \rho_p\sigma_p^2)W \tag{11}$$

$$= \sigma_a^2 \cdot (I_n + \rho_a \cdot W) \tag{12}$$

$$\rho_a = f\rho_d + (1-f)\rho_p,\ f = \sigma_d^2/\sigma_a^2,\ \sigma_a^2 = \sigma_d^2 + \sigma_p^2 \tag{13}$$

In this situation the overall level of dependence averages the individual levels of choice dependence by their relative variances. Of course, $\rho_d = \rho_p$ results in the conventional case. In addition, one can perform a similar analysis for

other spatial specifications such as SAR and CAR. These become slightly more complicated with averaging ρ_d^k and ρ_p^k for $k = 1 \cdots \infty$, but show some of the overall flavor of the simpler MA specification. This development highlights the possibilities created with different levels of choice dependence, choice variances, and spatial specifications[4].

3 The FBSGEV Choice Model

To model borrowers' choices, we compute the difference between the utilities of two choice alternatives d and p

$$\boldsymbol{U}_d - \boldsymbol{U}_p = X(\boldsymbol{\beta}_d - \boldsymbol{\beta}_p) + \boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p. \tag{14}$$

McFadden (1978) have assumed that the error term $\boldsymbol{\epsilon}_j$ is Gumbel distributed as the decision-maker's objective is to maximize his or her utility and the Gumbel distribution is used to model the distribution of the maximum (Embrechts et al. 2003). As the difference between two Gumbel random variables is a logistic distribution (Johnson et al. 2005), the difference $\boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p$ is assumed to be logistic distributed. The main limitation of this assumption is that it uses a symmetric distribution for the difference of the error terms $\boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p$.

There are several reasons supporting a skewed distribution for the error term $\boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p$. Firstly, if there is an omitted variable Z, the equation (14) becomes

$$U_d - U_p = X(\boldsymbol{\beta}_d - \boldsymbol{\beta}_p) + \boldsymbol{z}(\boldsymbol{\alpha}_d - \boldsymbol{\alpha}_p) + \boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p. \tag{15}$$

[4]We are indebted to a reviewer for suggesting this situation and this opens up possibilities for new spatial choice models.

This means that the error term $\boldsymbol{z}(\boldsymbol{\alpha}_d - \boldsymbol{\alpha}_p) + \boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p$ in the equation (15) is skewed distributed if the omitted variable $\boldsymbol{z}$ is also asymmetrically distributed. Secondly, even in the absence of omitted variables, if there is misspecification of the independent variables X, such as using a model with the explanatory variables in levels when they are in log-form, the error term $\boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p$ might have a skewed distribution. Thirdly, if ϵ_d and ϵ_p have symmetric distributions, but unequal variances, the disturbances $\boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p$ will have an asymmetric distribution.

To better understand the implications under different assumptions on the distribution of the error term $\boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p$, we perform a simulation study on three possible distributions for the disturbances given by the Normal, Logistic and the GEV. Table 1 shows that the GEV distribution gives more tail weight to large disturbances than their symmetric counterparts (Normal and Logistic). From a decision theory perspective, providing too little probability to rare events with large losses, such as default, has a higher error cost than providing too little probability to common events with small profits such as a loan repayment. Implicitly, using a skewed distribution combines the density and loss functions together.

For all the reasons stated above, we choose the GEV distribution to model the error term. As we point out in Section 2.1, we use a spatial binary choice model to take into account omitted variables in mortgage default decisions that are spatially dependent. This means that the equation (14) becomes

$$\boldsymbol{U}_d - \boldsymbol{U}_p = X(\boldsymbol{\beta}_d - \boldsymbol{\beta}_p) + A(\mathbf{v}_d - \mathbf{v}_p) \tag{16}$$

Distributions	1	2	3	4
Normal	0.1587	0.0228	0.0014	0.0000
Logistic	0.1402	0.0259	0.0043	0.0007
GEV(0)	0.1442	0.0423	0.0119	0.0033
GEV(0.10)	0.1301	0.0429	0.0154	0.0061
GEV(0.20)	0.1124	0.0400	0.0168	0.0080
GEV(0.25)	0.1021	0.0372	0.0165	0.0083
GEV(0.30)	0.0904	0.0335	0.0154	0.0082
GEV(0.35)	0.0776	0.0289	0.0138	0.0077
GEV(0.40)	0.0627	0.0232	0.0114	0.0065
GEV(0.45)	0.0477	0.0175	0.0088	0.0052

Table 1: Upper Tail Probability $P\{\epsilon > j \cdot \sigma\}$ with j = 1,2,3,4 and $E(\epsilon\epsilon') = \sigma^2 I_n$ for different distributions of the error term ϵ. The results are based on simulated data obtained from one billion standardized variates for each distribution.

where the error component is GEV distributed with a cumulative distribution function

$$F_{\text{GEV}}(v_{ij}) = \begin{cases} \exp\left\{ -\left[1 + \tau\left(\frac{v_{ij}-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\tau}} \right\} & \tau \neq 0 \\ \exp\left[-\left(\frac{v_{ij}-\mu}{\sigma}\right)\right] & \tau = 0 \end{cases} \tag{17}$$

where τ is the shape parameter, $\mu \in R$ is the location parameter, $\sigma \in R^+$ is the scale parameter and $x_+ = max(x,0)$. For simplicity, we consider $\mu = 0$ and $\sigma = 1$.

The GEV distribution is very flexible with the shape parameter τ controlling the tail behavior, as shown by Figure (1). Three groups of distributions are defined based on the value of the parameter τ. If $\tau \to 0$, the GEV dis-

tribution is the Gumbel class used by McFadden (1978) in the logit choice model. The distributions associated with $\tau > 0$ are called Fréchet-type distribution. Finally, in the case where $\tau < 0$, the distribution class is Weibull. To measure the skewness of the GEV distribution, we use the skewness measure proposed by Arnold and Groenveld (1995) as it only requires the existence of mode M. For a cumulative distribution function F, this skewness measure is $\gamma_M = 1 - 2F(M)$. The measure γ_M satisfies $-1 < \gamma_M < 1$, with 1(-1) indicating extreme right (left) skewness. For the GEV distribution (17), we obtain

$$\gamma_M = 1 - 2\exp[-(1+\tau)] \tag{18}$$

if $\tau > -1$; otherwise, the mode does not exist. The GEV distribution has negative skewness for $\tau < \ln(2) - 1$, it is positively skewed for $\tau > \ln(2) - 1$, and near symmetric for $\tau = \ln(2) - 1$.

Figure 1 around here

A logit choice model provides the same contribution to data on defaults ($Y_i = 1$) and non-distressed mortgages ($Y_i = 0$) (Calabrese et al. 2015; King and Zeng, 2001). As the two groups of borrowers are imbalanced with a lower percentage of defaults, an additional decision of default is more informative than a payment choice. Hence, we assign more weight to default choices using a GEV distribution instead of a logistic distribution. Moreover, the mortgage default decisions are represented by the tail of the utility function. The GEV random variable has been used in the literature to model the tail behavior (Embrechts et al. 2003). Another important advantage of the GEV distribu-

tion is that the marginal distributions of a multivariate GEV are also GEV distributed (Johnson et al. 2005).

We point out that the GEV model proposed by McFadden (1978), more properly defined as Gumbel model, suffers from the restriction of homoscedastic disturbances across choice alternatives. In a choice model, if heteroscedasticity across choices is ignored, the distribution of the error terms can be skewed (Yatchew and Griliches, 1985). Bhat (1995) proposed an extreme value model with heteroscedasticity across alternatives. This was further generalized by Zeng (2000), who developed a logit model with heteroscedasticity across decision makers as well as across alternatives. This model is also referred to as the heteroscedastic logit model (DeShazo and Fermo, 2002) and the parametrized heteroscedastic multinomial logit model (Hensher et al. 1999).

The homoscedasticity assumption across alternatives could be violated by mortgage default choice, as the decision of default may have a higher level of uncertainty than the choice of repayment. We remove the homoscedasticity assumption across alternatives. In particular, we assume that the ratio between the variance of the disturbances for payment v_{ip} and the variance of the error terms for the default alternative v_{id} is almost zero, $var(v_{ip})/var(v_{id}) \approx 0$. Under this assumption, the error component $\mathbf{v}_d - \mathbf{v}_p$ in the equation (16) is GEV distributed

$$\mathbf{v}_d - \mathbf{v}_p \sim GEV_n(\boldsymbol{\mu} = \mathbf{0}, I_n, \tau) \tag{19}$$

where I_n is the identity matrix. We define this model the Fast Binary Spatial GEV (FBSGEV) choice model. The FBSGEV model includes a near symmetric distribution for the error terms $\mathbf{v}_d - \mathbf{v}_p$ as a special case when $\tau = \ln(2) - 1$.

3.1 The estimation procedure suitable for large sample size

To estimate the FBSGEV model, given the equation (19), we have to compute the integral of a truncated n-dimensional GEV distribution

$$F_{n,\text{GEV}}(\mathbf{b}) = \int_{-\infty}^{b_n} \int_{-\infty}^{b_{n-1}} ... \int_{-\infty}^{b_1} f_{n,\text{GEV}}(v_1,v_2,...,v_n) dv_1 dv_2 ... dv_n \qquad (20)$$

where $v_i = v_{id} - v_{ip}$, $\mathbf{b} = [b_1,b_2,...,b_n]$, $f_{n,\text{GEV}}$ and $F_{n,\text{GEV}}$ are, respectively, the n-dimensional density and cumulative distribution function of a GEV random variable.

This becomes a more difficult computational problem as the sample size n increases. Different methods have been proposed for computing these integrals, such as the frequency simulator and the Stern simulator (Borsch-Supan and Hajivassiliou, 1993). A widely used technique is the smooth recursive conditioning simulator, known also as the Geweke-Hajivassiliou-Keene (GHK) simulator (Geweke, 1991; Hajivassiliou and McFadden, 1990; Keane, 1994). The GHK method reduces the integral of a truncated multivariate normal to a recursive sequence of n univariate integrals. Beron and Vijverberg (2004) have used the GHK method to propose the Recursive Importance Sampling (RIS) to estimate the parameters of a spatial probit. Using the Cholesky decomposition, they obtain a Cholesky triangular matrix of $n(n+1)/2$ non zero elements. This means that the RIS requires $O(n^2)$ operations to compute the multivariate integral, which becomes computationally intensive for large sample sizes (of the order of thousands of observations).

Beron and Vijverberg (2004) have proposed a RIS estimator to evaluate directly the n-dimensional integral. By using a decomposition of the

n-dimensional variance-covariance matrix that produces an upper-triangular matrix, the sampler can proceed by exploiting the fact that the last observation is now independent of other observations. The second-last observation is only dependent on the last, and so forth, thus allowing a recursive sampling algorithm. The RIS-normal simulator is identical to what is sometimes called the Geweke-Hajivassiliou-Keane (GHK) simulator (Borsch-Supan and Hajivassiliou, 1993).

In spatial econometrics an observation depends only on a low number of nearby observations. This means that the spatial weight matrix A in equation (4) may contain a large proportion of zeros, so the matrix is defined as being sparse. For example, if there are on average six neighbors for each observation, the proportion of non-zeros in W is almost equal to $6/n$. Pace and LeSage (2016) have suggested to use the GHK algorithm for a sparse inverse variance-covariance matrix, known as a precision matrix. In particular, the authors consider the CAR model defined in equation (5) and show that if the precision matrix $\Psi = \Sigma^{-1}$ is sparse, the variance-covariance matrix Σ

$$E[(\boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p)(\boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p)'] = E(\boldsymbol{\epsilon}_d \boldsymbol{\epsilon}_d') = \Sigma_{CAR} = (I_n - \rho W)^{-1} \tag{21}$$

is not sparse. Note, the first identity in equation (21) follows from the heteroscedasticity assumption across alternatives ($var(v_{id})/var(v_{ip}) \approx 0$) presented in the previous section.

An important property of the GEV distribution is that a multivariate GEV random variable has GEV marginal distributions (Kotz et al. 2005). We can use this property to simplify the integral (20) replacing the multivariate joint

density with the product of n conditional densities where each conditional density function depends only on prior variables in the sequence (22).

$$F_{n,\mathrm{GEV}}(\mathbf{b})=\int_{-\infty}^{b_n}\int_{-\infty}^{b_{n-1}}\cdots\int_{-\infty}^{b_1} f_{\mathrm{GEV}}(v_n)f_{\mathrm{GEV}}(v_{n-1}|v_{t>n-1})\ldots f_{\mathrm{GEV}}(v_1|v_{t>1})dv_1dv_2\ldots dv_n \quad (22)$$

We apply the GHK algorithm to a CAR model, defined in equation (21), and we compute the integral (22) using the Cholesky decomposition on the precision matrix, which results in a lower triangular matrix L and an upper triangular matrix Q, where Q is equal to the transpose of L ($Q = L'$). In particular, we aim to multiply $\mathbf{v} = \mathbf{v}_d - \mathbf{v}_d$, defined in equation (19), by a matrix to obtain the vector $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_d - \boldsymbol{\epsilon}_p$ of correlated GEV random variables whose variance-covariance matrix is Σ_{CAR}, given by equation (21). To achieve this aim, we consider the following equations

$$\begin{aligned} \Psi &= LQ = \Sigma^{-1} \\ \Sigma &= (LQ)^{-1} = Q^{-1}L^{-1} \\ \boldsymbol{\epsilon} &= Q^{-1}\mathbf{v} = L\mathbf{v} \\ E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') &= E(Q^{-1}\mathbf{v}\mathbf{v}'L^{-1}) = \Sigma. \end{aligned} \quad (23)$$

From equations (22) and (23), we obtain

$$Q\boldsymbol{\epsilon} = \mathbf{v} \text{ s.t. } \epsilon_i < b_i \text{ for } i = 1,2,\ldots,n.$$

$$\begin{bmatrix} v_1 \\ v_2 \\ ... \\ v_n \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{12} & ... & Q_{1n} \\ ... & & & \\ & & Q_{(n-1)(n-1)} & Q_{(n-1)n} \\ & & & Q_{nn} \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_n \end{bmatrix} \text{ s.t. } \begin{bmatrix} \epsilon_1 \\ ... \\ \epsilon_{n-1} \\ \epsilon_n \end{bmatrix} < \begin{bmatrix} b_1 \\ ... \\ b_{n-1} \\ b_n \end{bmatrix} \quad (24)$$

We point out that if the GHK algorithm is applied to the precision matrix as in (24), and not to the covariance matrix, the procedure begins with the last observation n and works towards the first observation. We can rewrite the system (24) in the following form

$$\begin{aligned} b_n &> \frac{v_n}{Q_{nn}} \\ b_{n-1} &> \frac{v_{n-1} - Q_{(n-1)n}\epsilon_n}{Q_{(n-1)(n-1)}} \\ &... \\ b_1 &> \frac{v_1 - \sum_{t=2}^{n} Q_{1t}\epsilon_t}{Q_{11}} \end{aligned}$$

The GHK procedure begins with the n-th observation that does not depend on any other observation, so the calculation of the n-th probability becomes an univariate problem

$$\begin{aligned} a_n &= b_n Q_{nn} \\ \overline{P}_n &= F_{\text{GEV}}[v_n < a_n] \\ v_n^* &\sim TGEV(a_n) \\ \epsilon_n^* &= \frac{v_n^*}{Q_{nn}} \end{aligned}$$

where $TGEV$ is a truncated GEV random variable. For the general i-th

observation, the ϵ_{i+1}^*, ..., ϵ_n^* calculated in the previous steps are used as follows

$$
\begin{aligned}
a_i^* &= b_i Q_{ii} + \sum_{t=i+1}^{n} Q_{it}\epsilon_t^* \\
\overline{P}_i &= F_{\text{GEV}}[v_i < a_i^*] \\
v_i^* &\sim TGEV(a_i^*) \qquad (25) \\
\epsilon_i^* &= \frac{v_i^* - \sum_{t=i+1}^{n} Q_{it}\epsilon_n^*}{Q_{it}}. \qquad (26)
\end{aligned}
$$

For the first observation (last in the process), v_1^* and ϵ_1^*, defined respectively in (25) and (26) do not need to be computed. If we repeat this procedure R times, we can follow Pace and LeSage (2016)'s proposal of computing the joint probability $\overline{P}$ as follows

$$
\overline{P} = \sum_{i=1}^{n} \ln\left(\frac{\sum_{d=1}^{R} \overline{P}_i(d)}{R}\right).
$$

We propose the previous procedure to estimate the unknown parameters of the FBSGEV model defined in equation (14) via a standard outer-product-of-the-gradient (OPG) method of optimization. We could use this procedure to estimate also the shape parameter τ of the GEV distribution (17). However, usual asymptotic properties associated with the maximum likelihood estimator are not satisfied when $\tau < -0.5$ (Smith, 1985). Hence, we propose fitting as many FBSGEV models as the number of a set of sensibly chosen values of the parameter τ and then select the model that yields the best empirical predictive performance.

4 Empirical analysis

4.1 Data

We selected Clark County, in the US state of Nevada, as the study area for this analysis because (a) it epitomized the mortgage crisis and (b) the Metropolitan Statistical Area (MSA) lies entirely in a single county. Therefore, we only need to obtain one county of property records to analyze a large city. This is in contrast to cities such as Denver, Colorado (10 counties); Charlotte, North Carolina (6 counties); or Dallas, Texas (12 counties). Las Vegas, the most populous city in Clark County, has the largest concentration of subprime mortgage origination in the country (Mayer and Pence, 2008), therefore it was hit hard by mortgage foreclosures and collapsing prices.

We collected information about individual homeownership and housing transactions from the Clark County Property Assessor's Office records. These records are comprised of three distinct data sets. The first file contains information on the physical characteristics of each single-family property located within Clark County such as the year the property was built, the square footage of the house, and the lot size. The second file contains transactions information for each of these properties. We can observe all transactions on a property between 2000-2011. Therefore, if we see that a mortgage has been originated in association with a property during this time period and there is no record of the loan being repaid during the same period, the loan will be included in our sample. Additionally, we can see information about the loan type (fixed or adjustable rate mortgage) and estimates on the current market value of the

property derived from tax assessment records.

The final data set contains information on mortgage default. The specific record we observe is the formal filing of a notice of default. This is a notice sent by mortgagee to the mortgagor when the borrower is 90 days or more delinquent in payment. The lender is not obligated to send this notice at the point of 90 days delinquency; the lender may rationally choose to offer a modification or some other loan workout to the borrower. However, sending this legal notice is a necessary precursor for the lender to initiate foreclosure proceedings. We can observe the property each notice is associated with as well as the date that the notice was sent. Collectively, these records include information on property transactions for every single family property in Clark County. Default records include the date of each notice of default filed against each property in Clark County. Using this information, we can ascertain if an individual received a notice of default during the relevant time period or not. We use this information as the dependent variable in the empirical specifications predicting default.

We only include property sales records from individuals owning residential property. If the owner's name included any word indicating a business, we excluded the associated record. Those words include: LLC, Inc, Residential, Property, Properties, Construction, Finance, Resort, Vacation, Mortgage, Financial, Global, Bank, Home, Security, Securities, Services, Servicing, Nevada, Fund, Wells Fargo, Consultant, or Series. Properties that do not have a mortgage in place at the time of analysis are excluded from the sample. In addition, we required observations to have loan-to-value ratios between 0 and 4, to have

only a senior mortgage loan, to have complete data on the interest rate type and on the location of the property.[5]

The final data set included 282,366 observations. We look to variables from the 2010 American Community Survey (ACS) estimates for Clark County to ascertain whether this number of observations is in line with the true population. According to the ACS, there were 303,652 owner-occupied housing units with a mortgage. Of this number, 24,737 are reported to have a second mortgage. The different between these two measures yields 278,915 observations; given that the stated margin of error is $+/-$ 7,730 for the number of mortgages and $+/-$ 2,377 for the number of second mortgages, we conclude that the data set we construct from Assessor's Office records is well in line with the true number of owner-occupied properties with a mortgage, but no second mortgage.[6]

We code payments as ones and default as zeros. We estimate a model for 2009 and one for 2010 to analyze the performance of the FBSGEV model for different percentages of default. If we observe the dependent variable in 2009, the default rate is 2.7%. We added the defaulted mortgage loans in 2010 to the defaults in 2009, therefore the default rate in 2010 increases to 5.54%.

The Las Vegas borrowers preferred fixed rate loans (82.1%) as opposed to adjustable rate loans (17.9%). In 2009-10 most of the borrowers already owed more on their mortgages than the estimated market value of their house. The median loan-to-value ratio was 1.038, and 22.5% of the borrowers had equity

[5]Observations with a recorded second mortgage are excluded from the sample.

[6]Additionally, 2010 Census reports that Clark County has a population of 1,951,269 persons organized into 735,475 households. Combined with ACS data from 2010 stating that approximately 70% of housing units in Clark County in 2010 have some form of mortgage associated with the property and additionally, the home ownership rate for Clark County, 59%, we get to approximately the same number of mortgage observations.

positions of one-third or less relative to their obligation. Nonetheless, the rate of observed delinquency was not as high one might expect. This is largely due to the stringent definition of measure of delinquency we use.

To summarize our prior discussions of the model and to give the actual empirical specification, we restate the model assumptions in (27) through (31). Specifically, in (27) we posit that the latent utility difference between default and payment for individual i depends on $\ln(L/T)$, the logarithm of loan-to-value, and FR, a dummy variable for a fixed rate mortgage as well as the difference in the disturbances associated with the default and payment choices. In (28) and (29) we associate the observed binary choice Y_i with the latent utility difference ΔU_i. In (30) we assume the marginal distribution of the disturbances follows a Generalized Extreme Value distribution with an expected value of 0 and a scale of 1. As discussed in Train (2009), for identification binary choice models typically impose a fixed scale as in (30). Finally, in (31) A specifies spatial dependence as a function of a parameter $\rho \in [0,1)$.

From the equation (2), we estimate the following model

$$\Delta U_i = \beta_1 + \ln(L/T) \cdot \beta_2 + \mathrm{FR} \cdot \beta_3 + A(\rho)\Delta\epsilon_i \tag{27}$$

$$\Delta U_i > 0 \rightarrow Y_i = 1 \tag{28}$$

$$\Delta U_i < 0 \rightarrow Y_i = 0 \tag{29}$$

$$\Delta\epsilon \sim GEV(\tau,0,1) \tag{30}$$

$$A(\rho) = (I_n - \rho W)^{-1/2} \tag{31}$$

We also address matched the observations to obtain locational coordinates

used to create the spatial weight matrix W. In all our analysis we use a Delaunay triangle routine to determine a contiguity-based W (the most common in the literature, see for example LeSage and Pace, 2009) and standardize it so that the rows and columns sum to 1. The resulting W contains non-negative elements when observations i and j neighbor each other. Following the literature, we do not allow observations to neighbor themselves and thus set $w_{ii} = 0$ for $i \dots n$. Therefore, W is a non-negative, doubly stochastic, and symmetric matrix where $\text{tr}(W) = 0$.

4.2 Empirical results

We examine the performance of a number of binary choice models on the Las Vegas mortgage data for different percentages of borrowers who defaulted on their mortgage loans. In 2009 the percentage of default is 2.7%, in 2009-2010 it increases to 5.54%. The dependent variable Y is coded as 1 if the borrower decided to repay his/her mortgage loan, 0 otherwise.

The tail of the response curve for values close to 0 represents features. Hence, a positive skewed GEV distribution is more suitable for an imbalanced binary sample with a low percentage of zeros, such as in this empirical analysis. Otherwise, a negative skewness is preferred if one represents the rare event. For this reason, we choose values of the parameter τ such that the GEV distribution is positive skewed ($\tau = 0.45, 0.35, 0.30, 0.25, 0.20$). We compare the performance of FBSGEV(τ) with those of spatial probit, independent and identically distributed (*iid*) probit, logit, and loglog estimators. Table 2 shows the estimates.

The alternative with the highest utility is the same no matter how utility is scaled. Therefore, adding a constant to the utility of all the alternatives or multiplying each borrower's utility by a constant does not change the decision maker's choice. To take account of this, the scale of utility must be normalized, so that it is equivalent to normalizing the variance of the error terms (Train, 2009). As the variance of the disturbances changes in the models analyzed in Table 2, the estimated parameters have an arbitrary identification. To deal with the identification issue, one can focus on the t values or on the relative parameter estimates as in Table 3. In contrast, the spatial parameter is identified. The various spatial GEV estimates show similarities in both their estimated t values and in their relative parameter estimates. In Table 3 we record the change in the log-likelihood from the estimator giving the highest log-likelihood value (FBSGEV(0.40)) in column ΔL. Finally, we give the running times of the various estimators in terms of minutes in the column labeled Time.

Table 2 exhibits some trends across choice models. First, the spatial estimators show greater precision for the non-constant explanatory variables, with the log of the loan-to-value ratio $t_{L/V}$ showing material changes from around -36.72 for the *iid* GEV model with the highest likelihood to -49.48 for the highest likelihood spatial estimator (FBSGEV(0.40)). The literature on mortgage defaults has widely recognized the loan-to-value ratio as one of the most important determinants of borrowers' decisions (Garmaise, 2015; Elul, 2016; Lin, 2014; Kau et al. 2014). The ρ parameter estimates range from a high of 0.484 for spatial probit to 0.389 for FBSGEV(0.45). The ρ and τ param-

	$\tilde{\beta}_C$	$\tilde{\beta}_F$	$\tilde{\beta}_{L/V}$	$\tilde{\rho}$
Probit	1.738	0.269	−0.324	
	153.645	20.829	−33.450	
Logit	3.204	0.606	−0.885	
	125.377	20.993	−36.340	
GEV(0)	3.229	0.594	−0.877	
	129.536	21.088	−36.719	
SProbit	1.793	0.279	−0.340	0.485
	147.440	22.755	−54.417	20.945
SGEV(0.45)	9.155	2.126	−4.692	0.388
	86.014	21.258	−49.148	31.738
SGEV(0.40)	8.135	1.851	−4.024	0.409
	89.948	21.539	−49.664	31.665
SGEV(0.35)	7.243	1.620	−3.433	0.432
	92.640	21.870	−50.226	29.057
SGEV(0.30)	6.468	1.414	−2.907	0.448
	95.269	22.174	−51.029	27.009
SGEV(0.25)	5.794	1.234	−2.442	0.461
	97.812	22.459	−51.790	25.589
SGEV(0.20)	5.204	1.076	−2.032	0.470
	99.684	22.713	−52.470	24.090

Table 2: Estimate of Probability of Payment ($Y = 1$) Across Estimators Based on the Observations in 2009 (the Default Rate is 2.7%).

	$\tilde{\beta}_F/\tilde{\beta}_C$	$\tilde{\beta}_{L/V}/\tilde{\beta}_C$	$\tilde{\beta}_F/\tilde{\beta}_{L/V}$	ΔL	Time
Probit	0.155	−0.186	−0.829	−419.919	0.009
Logit	0.189	−0.276	−0.685	−303.462	0.004
GEV(0)	0.184	−0.272	−0.676	−294.201	0.004
SProbit	0.156	−0.190	−0.820	−287.932	20.764
SGEV(0.45)	0.232	−0.512	−0.453	−0.924	11.615
SGEV(0.40)	0.228	−0.495	−0.460	0.000	10.121
SGEV(0.35)	0.224	−0.474	−0.472	−5.824	10.141
SGEV(0.30)	0.219	−0.449	−0.486	−19.535	8.686
SGEV(0.25)	0.213	−0.421	−0.505	−39.382	10.068
SGEV(0.20)	0.207	−0.391	−0.530	−64.707	25.943

Table 3: Relative Estimates of Probability of Payment ($Y = 1$) , Difference in Log-Likelihood, and Timing (Minutes) Across Estimators Based on the Observations in 2009 (the Default Rate is 2.7%).

eter vary inversely. The t statistic for ρ always exceeds that for the fixed rate dummy for all the FBSGEV estimators, which indicates its importance in terms of the fit. All of the *iid* choice models show a lower likelihood than the spatial methods, and all the FBSGEV models show a higher likelihood. The difference between the log-likelihoods of the spatial probit and the FB-SGEV(0.40) is large and statistically significant (283.658). Furthermore, the τ parameter makes a large difference in the log-likelihood. For example, the FBSGEV(0.40) shows a log-likelihood that is 60.35 above the FBSGEV(0.20). These results show the advantage of using a flexible asymmetric distribution with fat tails as the GEV distribution for the error terms.

Naturally, the *iid* choice models have trivial running times. However, the running times of the spatial models seem quite reasonable given the repeated computation of a 282,366 dimensional integral. Due to quicker convergence,

the running times typically fall as the log-likelihood rises. In fact, the estimation time for FBSGEV(40) is less than half the running time for the spatial probit. We do not estimate the FBSGEV(0) because of poor convergence for these very imbalanced data.

4.3 Credit risk assessment

Figure 2 around here

Financial institutions use binary choice models to classify potential default decisions. We compare the distributions of the estimated probability of repayment obtained under the different choice models. We show them in the box plots in Figure 2. By the nature of binary models, the estimated probabilities of repayment have the same means. However, the distribution in the tails varies substantially from *iid* probit, with a minimum estimated probability of around 0.9, to FBSGEV(0.45) with a minimum estimated probability of repayment around 0.25. Although risk managers can understand that the estimated probabilities from models may show less variation than the true probabilities, a naive user might interpret the *iid* estimated probabilities as indicating that there was little scope for default, which would have been the wrong conclusion for Las Vegas during the financial crisis. Introducing neighborhood effects in a probit model slightly increases the range of the estimated probabilities (the minimum of the probability of repayment reaches 0.75). Among the estimators considered herein only the FBSGEV approach can accurately model the left tail behaviors of the estimates, which is a crucial issue in the risk management of a mortgage portfolio as the left tails represent the defaulted borrowers.

Figure 3 around here

Misclassifying a defaulted mortgage loan (rare event) as a performing loan (common event) represents the most expensive form of error for the risk management of a credit portfolio. Figure 3 shows the performance of Probit, GEV(0)[7], spatial Probit, and FBSGEV(0.45) in terms of this form of misclassification as the rejection rate increases. For rejecting a small percentage of loans (less than 5%), the two *iid* models, probit and GEV(0), outperformed the spatial models by about one percent or less. However, as the percentage of rejected loans increases, the FBSGEV(0.45) begins to dominate the *iid* models.

According to an analysis conducted by Timiraos and Tamman (2011) on mortgage data filed with banking regulators, the 10 largest mortgage lenders in the US denied 23.5% of mortgage applications in 2009 and 26.8% in 2010. Therefore, if these observations represented potential applicants, at a rejection rate of 25%, FBSGEV(0.45) shows more than a 4% improvement in misclassification of defaulted mortgages relative to the probit models (spatial and *iid* methods) and more than a 2% improvement relative to the *iid* cloglog model. As the rejection rate increases from 30%, the performance of the cloglog model is poorer and the non-spatial methods do worse than the spatial approaches. In this range, FBSGEV(0.45) shows around a 6% improvement in misclassification of the rare event relative to the *iid* models. At a 99% rejection rate (cherry picking the top 1% of the highest ranked loans for each method), the FBSGEV(0.45) does very well with a performance improvement of over 8% in misclassification relative to the *iid* methods.

[7]The GEV(0) corresponds to the cloglog model (Agresti, 2002).

To analyze the performance of FBSGEV as the percentage of the rare event changes in the sample, we examine the loans over the period of 2009-10 where the average of Y equals 0.9446, or a default rate of 5.54%. Table 4 displays that, also in this case, the FBSGEV models shows a higher log-likelihood. Particularly, FBSGEV(0.35) shows the highest log-likelihood, while *iid* probit shows the smallest log-likelihood. From equation (18), we compute the skewness measure proposed by Arnold and Groeneveld (1995). We obtain 0.4815 for FBSGEV(0.35) and 0.5068 for FBSGEV(0.40). In line with expectations, with a less imbalanced Y the skewness of the FBSGEV model with the best performance decreases.

In contrast with the previous results for 2009 data, the *iid* logit and GEV(0) show higher log-likelihoods than the spatial probit model. The estimation times in Table 4 decrease relative to those in Table 2, consistent with the added information coming from a less imbalanced binary sample. Also in this case, the estimation time for the FBSGEV(0.35) models is very low in comparison with the time for the spatial probit, the first is less than half of the latter. There is still an inverse relationship between the parameters ρ and τ. The estimates of ρ are higher than those shown in Table 2, preserving their ordering. In line with the results in 2009, the t statistic for ρ exceeds that for the fixed rate dummy for all the FBSGEV models. Instead, the spatial probit model shows a lower t statistic for ρ than that for the fixed rate dummy.

Figure 4 around here

We represent the estimated probabilities of mortgage payment in the box plots in Figure 4. In line with the previous results, the *iid* models show a

	t_C	t_F	$t_{L/V}$	ρ	t_ρ	ΔL	Time
Probit	153.957	25.771	−48.846			685.940	0.008
Logit	134.037	25.679	−53.005			438.106	0.003
GEV(0)	143.676	25.840	−54.103			401.460	0.004
SProbit	157.201	28.056	−78.443	0.437	25.784	501.833	18.790
FBSGEV(0.45)	111.763	25.313	−72.719	0.348	35.178	19.791	8.870
FBSGEV(0.40)	114.758	25.788	−72.805	0.368	35.969	2.084	7.360
FBSGEV(0.35)	117.184	26.175	−73.045	0.384	34.732	0.000	7.346
FBSGEV(0.30)	119.386	26.594	−73.643	0.398	33.550	11.777	8.827
FBSGEV(0.25)	121.809	26.966	−74.532	0.412	32.795	35.819	11.792
FBSGEV(0.20)	123.788	27.332	−75.458	0.424	31.940	74.999	17.647

Table 4: Estimate of Probability of Payment ($y = 1$) Across Estimators based on the Observations in 2009-10 (the default rate is 5.54%).

lower variability of the estimated probabilities of loan payment than those obtained by the FBSGEV models. For example, the minimum probability of repayment for the *iid* probit model is around 0.8. Therefore, ignoring spatial dependence could be the cause of incorrect decisions when assessing default risk for mortgage borrowers. Even if we include the neighborhood effects in a probit model, the minimum estimated probability of being a performing loan is about 0.6. This result is in line with those obtained in a non-spatial framework by King and Zeng (2001), Wang and Dey (2010) and Calabrese et al. (2015). Probit and logit have relatively thin tails and for imbalanced samples with rare events, these models do not assign much probability to the rare event. Credit standards lead to this imbalance and therefore logit and probit may experience difficulties when dealing with mortgage data.

Figure 5 around here

Figure 5 shows the performance of Probit, FBSGEV(0), spatial Probit, and FBSGEV(0.45) in terms of the misclassification of default as the rejection rate increases. For rejecting a small percentage of loans (less than 5%), the two *iid* models, probit and GEV(0), outperform the spatial models by about 1% or less. However, as the percentage of rejected loans goes over 5%, the FBSGEV begins to dominate the *iid* approaches and the spatial probit model. At a rejection rate of 25% (Timirao and Tamman, 2011), the FBSGEV(0.45) shows about a 3% improvement in misclassification of a rare event relative to the *iid* and spatial probit models. The improvement increases at 5% at a rejection rate of 50%. As the rejection rate increases from 40%, spatial probit begins to improve and actually does about the same at around 90% as FBSGEV(0.45). At the 99% rejection (cherry picking the top 1% of the highest ranked loans for each method), the FBSGEV(0.45) does very well with a performance improvement of over 6% in misclassification relative to the *iid* methods.

As Value-at-Risk (VaR) is the official measure of credit risk and constitutes the central point to the determination of capital requirements (Basel Committee on Banking Supervision, 2010), we compute the VaR for the Loss Distribution for different confidence levels. Basel II and III guidelines establish that the loss is the product between the Probability of Default (PD), Loss Given Default (LGD) and Exposure at Default (EAD) $Loss = PD \cdot LGD \cdot EAD$. As we do not have any information about the LGD and our aim is to perform a comparative analysis, we consider the loss as the product between PD and the loan value. Therefore, we assume that LGD is constant proportion for all

Confidence level	0.95	0.99	0.999
Models		$1 < LTV < 2$	
Probit	13,797.431	23,276.440	38,100.510
Logit	14,402.015	25,310.349	43,658.394
GEV(0)	14,442.268	25,390.815	44,150.333
SProbit	14,043.111	24,116.328	44,101.706
FBSGEV(0.45)	13,663.965	26,353.052	61,490.107
FBSGEV(0.40)	13,938.036	27,117.590	62,425.963
FBSGEV(0.35)	14,211.432	27,604.684	62,805.543
FBSGEV(0.30)	14,386.410	27,583.467	62,863.133
FBSGEV(0.25)	14,564.271	27,742.540	60,610.901
FBSGEV(0.20)	14,711.730	27,511.061	59,199.712
Models		$LTV > 2$	
Probit	13,282.077	22,041.712	36,653.456
Logit	13,918.716	23,979.318	41,194.576
GEV(0)	13,983.279	24,224.738	42,096.184
SProbit	13,640.904	22,492.317	41,389.561
FBSGEV(0.45)	13,266.947	25,386.768	54,613.227
FBSGEV(0.40)	13,525.560	25,947.909	55,302.474
FBSGEV(0.35)	13,833.506	26,335.914	55,723.940
FBSGEV(0.30)	14,014.923	26,296.526	55,375.011
FBSGEV(0.25)	14,145.043	26,269.438	54,623.861
FBSGEV(0.20)	14,239.471	26,393.733	55,058.068

Table 5: VaR for the Loss Distribution Across Models on Data in 2009 For Different Levels of Confidence and Buckets Based on loan-to-value (LTV).

loans in a given loan-to-value range, analogous to the standardized internal ratings based (IRB) approach under the Basel II guidelines (Basel Committee on Banking Supervision, 2005). We compute the VaR of the loss distribution at different levels of confidence and loan-to-value (LTV) on the data observed in 2009. We report the results in Table 5. Within each range of loan-to-value ratios, the difference between the VaR computed under independence and a symmetric distribution increases as the level of confidence increases. As the capital requirements are based on such estimates, this implies that financial institutions could underestimate the levels of credit risk on their portfolios and, hence, their regulatory capital.

We illustrate this further with some figures from Table 5. If the LGD is a constant proportion for all loans in a given range of loan-to-value, it will cancel when examining a ratio of two estimates of VaR. So for loans in 2009 having loan-to-value ratios between 1 and 2 and with a confidence of 0.999, logit would yield an estimated VaR of \$43,658 while the highest likelihood FBSGEV with a $\tau = 0.4$ would yield an estimated VaR of \$62,425. Therefore, the highest performing FBSGEV with $\tau = 0.4$ had a 43% higher VaR estimate than logit.

5 Robustness Checks

In this section we examine variations in the base or reference regression in Table 4. Specfically, we look at the effects of changing the number of random replications R as well as the type (pseudo-random, quasi-random), different weight matrices, and expanding the number of explanatory variables. We begin

with Table 6 where in Panel A we examine using R equal to 25,50,100, and 200 in the calculation of the GHK using a uniform random number generator (rand in Matlab 2017b). In Panel A the estimate of ρ rises from 0.381 to 0.392 as R goes from 25 to 100, but declines to 0.391 for $R = 200$. The range of estimates equals 0.011 which lies below the statistical noise in the estimation of ρ. Various authors have suggested using quasi-random numbers as an improvement over pseudo-random numbers in the GHK. We examine quasi-random numbers using Halton and Sobol sets in Panels B and C.[8] These show very little change from using the typical pseudo-random numbers in this application.

[8]We used the example setting for these from the Matlab commands haltonset and sobolset.

R	$\tilde{\beta}_C$	$\tilde{\beta}_F$	$\tilde{\beta}_{L/V}$	$\tilde{\rho}$	Time
Panel A: Pseudo-Random Uniform					
25	5.169	1.058	−2.644	0.381	4.214
50	5.176	1.055	−2.654	0.387	5.341
100	5.184	1.054	−2.664	0.392	7.203
200	5.183	1.054	−2.662	0.391	10.790
Panel B: Halton Quasi-Random					
25	5.178	1.051	−2.656	0.381	4.222
50	5.174	1.055	−2.659	0.389	5.348
100	5.179	1.056	−2.662	0.391	7.161
200	5.181	1.055	−2.662	0.390	10.832
Panel C: Sobol Quasi-Random					
25	5.171	1.057	−2.650	0.379	4.206
50	5.179	1.055	−2.660	0.386	5.348
100	5.187	1.051	−2.664	0.388	7.174
200	5.184	1.054	−2.664	0.390	10.851

Table 6: Variation in Estimates by Number of Repetitions R and Type (Pseudo-Random, Quasi-Random)

In Table 7 we examine using nearest neighbor weight matrices instead on the contiguity weight matrix. Specifically, we look at 4, 6,8, 10, and 12 nearest neighbors. We see similar estimates of β for all the specifications, although the likelihood rises with the number of neighbors and so does $\tilde{\rho}$ as well as the calculation time. The highest log-likelihood was for 12 nearest neighbors ($L = -57{,}216.37$), but this likelihood was less than the log-likelihood of the contiguity W ($L = -57{,}156.58$).

m	$\tilde{\beta}_C$	$\tilde{\beta}_F$	$\tilde{\beta}_{L/V}$	$\tilde{\rho}$	ΔL	Time
4	5.188	1.055	−2.586	0.353	−61.521	7.596
6	5.159	1.057	−2.650	0.375	−13.710	10.229
8	5.150	1.039	−2.672	0.386	−12.673	10.441
10	5.140	1.040	−2.696	0.394	−1.412	21.221
12	5.141	1.027	−2.709	0.401	0.000	51.262

Table 7: Estimates by W with Different Number of Nearest Neighbors

Finally, in Table 8 we explored the effects of adding other explanatory variables. Specifically, we added Age, Marital Status, and Gender, spatial lags of these variables as well as the FRM and logged L/V ratio, and a five degree polynomial in terms of the locational coordinates. Although Age, Marital Status, and Gender could have various effects, one possible channel to have an influence of mortgage behavior is through wealth and these variables are associated with wealth. Specifically, older couples tend to have the highest levels of wealth and younger, singles tend to have the lowest levels of wealth. Wealth provides the wherewithal to pay a loan, but also exposes the borrower to the potential to pay delinquency judgments should they decide to default. Therefore, wealth tends to lower the propensity to default in multiple ways. We see that age tends to increase the propensity to pay and that being single tends to reduce the propensity to pay. In terms of the spatially lagged explanatory variables all of these have the same signs as the individual variables and so the indirect effects reinforce the direct effects. As typical, these have lower levels of precision than the individuals direct effects. The fixed locational effects from the five degree polynomial and the constant term do not appear particularly important given the sample size and the parameter estimates from the expanded regression are comparable in magnitude to the estimates from the base regression. In particular, the $\tilde{\rho}$ is 0.389 with a t value of 35.42 in the base regression and is 0.399 with a t value of 36.49 in the expanded regression.

	$\tilde{\beta}$	t	$\tilde{\beta}_{\text{base}}$	t_{base}
FRM	1.060	26.923	1.052	26.173
$\ln L/V$	−2.323	−63.504	−2.651	−75.781
Age	0.816	16.866		
Single	−0.695	−17.729		
Female	0.343	7.782		
$W \cdot$ FRM	0.224	2.112		
$W \cdot \ln L/V$	−0.208	−4.070		
$W \cdot$ Age	0.368	3.456		
$W \cdot$ Single	−0.077	−0.900		
$W \cdot$ Female	0.231	2.462		
p_1	66.204	4.586		
p_2	−35.790	−1.987		
p_3	5.394	0.437		
p_4	−33.222	−2.617		
p_5	20.393	1.403		
Constant	0.529	1.181	5.179	117.379
$\tilde{\rho}$	0.399	36.487	0.389	35.421
L	−56,687.954		−57,156.580	
τ	0.35		0.35	
n	282,366			
Time (mins)	55.83		7.111	

Table 8: Estimate of Probability of Payment ($y = 1$) Across Estimators based on the Observations in 2009-10 (the default rate is 5.54%) Using an Expanded Model.

6 Conclusion

We introduced a spatial choice model that is accurate in classifying binary rare events and can handle large sample sizes. The proposed approach is based on a skewed and flexible distribution of the error terms, given by the GEV random variable. The tail behavior of the error distribution is determined by the rarity of the event in the sample, i.e. higher imbalanced samples are associated with higher skewness of the error distribution. If the dependent variable at each spatial location is binary, but the underlying latent variable is continuous, evaluating the likelihood function involves the integral of a truncated multivariate distribution of a dimension equal to the sample size. For large sample sizes, this becomes a difficult computational problem. Fortunately, each observation located in space may depend upon a small number of neighbors. This implies that the inverse of the variance-covariance matrix, known as a precision matrix, could be sparse. We exploit this sparsity by applying the Cholesky decomposition to the precision matrix. Therefore, we propose a variant of the Geweke-Hajivassiliou-Keane (GHK) algorithm, obtaining a number of computations almost linearly with the sample size ($O(n)$). Instead, spatial probit models using non-sparse methods based on the GHK algorithm require at least $O(n^2)$ computations. We define the suggested approach as the Fast Binary Spatial Generalized Extreme Value (FBSGEV) model.

Our proposal and its competitors were applied to data on 282,366 mortgages from 2009-2010 in Clark County, one of the areas with the largest concentration of subprime mortgages in the US. The empirical results confirmed that the main advantage of the FBSGEV model lies in its superior performance in

classifying potentially defaulted mortgages for different default rates in the sample. Another strength of this approach is that it provides more reliable estimates of the probabilities of repayment compared to classic alternatives. The empirical analysis also shows that spatial dependence had an important impact on model fit, as the t statistic for the spatial dependence parameter exceeded the t statistic associated with the fixed rate dummy.

The adoption of the FBSGEV model to analyze mortgage decisions can lead to some significant insights. Conventional models that ignore neighborhood effects can overestimate the probability of mortgage repayment. This is because a borrower has a higher propensity to repay, holding other things constant, when her/his neighbors also have a high propensity to repay. Therefore, the FBSGEV model can improve the internal assessments of financial institutions when they are evaluating mortgage decisions. It can also provide accurate evaluations of risk generated by relaxing mortgage underwriting standards, which occurred during the 2008/2009 financial crisis.

References

Agarwal, S., Amromin, G., Ben-David, I., Chomsisengphet, S., Piskorski, T. and Seru, A. (2012) Policy Intervention in Debt Renegotiation: Evidence from Home Affordable Modification Program, NBER Working Paper 18311.

Agresti, A. (2002) *Categorical Data Analysis*. Wiley.

Arnold, B. C. and Groeneveld, R. A. (1995) Measuring Skewness with Respect to the Mode. *The American Statistician*, 49, 34-38.

Basel Committee on Banking Supervision (2005) Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework. Bank for International Settlements.

Basel Committee on Banking Supervision (2010) Basel III: A global regulatory framework for more resilient banks and banking systems. Bank for International Settlements.

Beron, K.J. and Vijverberg, W.P.M. (2004) Probit in a Spatial Context: A Monte Carlo Analysis. In L. Anselin, R.J.G.M. Florax, and S.J. Rey, eds., *Advances in Spatial Econometrics: Methodology, Tools and Applications*. 62, 169-195, Springer-Verlag: Berlin.

Bhat, C. R. (1995) A Heteroscedastic Extreme Value Model of Intercity Travel Mode Choice. *Transportation Research*, 29, 471-83.

Borsch-Supan, A. and Hajivassiliou, V. A. (1993) Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of Econometrics*, 58(3), 347-368.

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.

Calabrese, R. and Elkink, J. A. (2016) Estimating Binary Spatial Autoregressive Models for Rare Events. *Advances in Econometrics. Spatial Econometrics : Qualitative and Limited Dependent Variables*, 145-166.

Calabrese, R,, Marra, G. and Osmetti, S. A. (2015) Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. *Journal of the Operational Research Society*, 67 (4), 604-615.

Deng, Y., Quigley, J.M. and Van Order, R. (2000) Mortgage terminations, heterogeneity and the exercise of mortgage options. *Econometrica*, 68(2), 275-307.

DeShazo, J.R. and Fermo, G. (2002). Designing choice sets for stated preference methods: the effects of complexity on choice consistency. *Journal of Environmental Economics and Management* 44, 123-143.

Elul, R. (2016) Securitization and Mortgage Default. *Journal of Financial Services Research*, 49 (2), 281-309.

Embrechts, P., Kluppelberg, C. and Mikosch, T. (2003) *Modeling Extremal Events for Insurance and Finance* Springer-Verlag.

Garmaise, M. J. (2015) Borrower Misreporting and Loan Performance. *Journal of Finance*, 70(1), 449-484.

Geweke J. (1991) Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints. *Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface. American Statistical Association*, 571-578.

Ghent, A. C. and Kudlyak, M. (2011) Recourse and Residential Mortgage Default: Evidence from US States. *The Review of Financial Studies* 24 (9), 3139-3186.

Hajivassiliou, V. and McFadden, D. (1990) The Method of Simulated Scores for the Estimation of LDV Models with an Application to External Debt Crises. *Cowles Foundation Discussion*, Paper 967, Yale University.

Harding, J., Rosenblatt, E. and Yao, V. (2009) The Contagion Effect of Foreclosed Properties. *Journal of Urban Economics*, 66(3): 164-78.

Hensher, D., Louviere, J. and Swait, J. (1999) Combining sources of preference data. *Journal of Econometrics* 89, 197-221.

Johnson, N. L., Kemp, A. W. and Kotz, S. (2005) *Univariate Discrete Distributions.* Third Edition, John Wiley & Sons, Inc.

Keane, M. (1994) A Computationally Practical Simulation Estimator for Panel Data. *Econometrica*, 62, 95-116.

Kau, J. B., Keenan, D. C. and Lyubimov, C. (2014) First Mortgages, Second Mortgages, and Their Default. *The Journal of Real Estate Finance and Economics*, 48, 4, 561-588.

King, G. and Zeng, L. (2001) Logistic Regression in Rare Events Data. *Political Analysis*, 9, 321-354.

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 and Economic Statistics*, 26(4), 460-471.

Kotz, S. Balakrishnan, N. and Johnson N. (2005) *Continuous Multivariate Distributions: Models and Applications.* Second Edition, John Wiley & Sons, Inc.

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

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

LeSage, J.P. and Pace, R. K. (2004) Models for spatially dependent missing data. *The Journal of Real Estate Finance and Economics*, 29, 233-254.

Lin, L. (2014) Optimal lean-to-value ratio and the efficiency gains of default. *Annals of Finance*, 10(1), 47-69.

Luce, D. (1959) *Individual Choice Behavior*, John Wiley and Sons, New York.

Mian, A., Sufi, A. and Trebbia, F. (2010) The Political Economy of the US Mortgage Default Crisis. *The American Economic Review* 10 (5), 1967-1998.

Mayer, C. and Pence, K. (2008) Subprime mortgages: What, where, and to whom? *Federal Reserve Board.* Washington D.C. Finance and Economics Discussion Series.

Mayer, C. and Pence, K., Sherlund, S. (2009) The Rise in Mortgage Defaults. *Journal of Economic Perspectives* 23, 27-50.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models.* Chapman Hall, New York.

McFadden, D. (1974) Conditional logit analysis of qualitative choice behavior, in P.Zarembka, ed., *Frontiers in Econometrics*, Academic Press, New York, pp. 105-142.

McFadden, D. (1978) Modeling the Choice of Residential Location. In *Spatial Interaction Theory and Planning Models*, edited by A. Karlqvist. Amsterdam: North-Holland.

Pace, R. K. and LeSage, J. P. (2016) Fast Simulated Maximum Likelihood Estimation of the Spatial Probit Model Capable of Handling Large Samples. *Advances in Econometrics. Spatial Econometrics : Qualitative and Limited Dependent Variables*, 3-34.

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

Sahare, M. and Gupta, H. (2012) A Review of Multi-Class Classification for Imbalanced Data. *International Journal of Advanced Computer Research*, 2(3), 2277-7970.

Scharlemann, T. C. and Shore, S. H. (2016) The effect of Negative Equity on Mortgage Default: Evidence From HAMP's Principal Reduction Alternative. *Review of Financial Studies*, 29(10), 2850-2883.

Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, 72, 67-90.

Timiraos, N. and Tamman, M. (2011) The Tighter Lending Crimps Housing. *Wall Street Journal*, June 25.

Train K. E. (2009) *Discrete Choice Methods with Simulation*. Second Edition, Cambridge.

Wang, X. and Dey, D. K. (2010) Generalized Extreme Value regression for binary response data: An application to B2B electronic payments system adoption. *Annals of Applied Statistics*, 4(4), 2000-2023.

Yatchew, A. and Griliches, Z. (1985). Specification Error in Probit Models. *Review of Economics and Statistics*, 67, 134-39.

Zeng, L. (2000). A Heteroscedastic Generalized Extreme Value Discrete Choice Model, *Sociological Method & Research*, 29(1), 118-144.

Zhu, S. and Pace, K. (2014) Modeling Spatially Interdependent Mortgage Decisions. *The Journal of Real Estate Finance and Economics*, 49(4), 598-620.

Zhu, S. and Pace, K. (2015) The Influence of Foreclosure Delays on Borrowers' Default Behavior. *Journal of Money, Credit and Banking*, 47 (6), 1205-1222

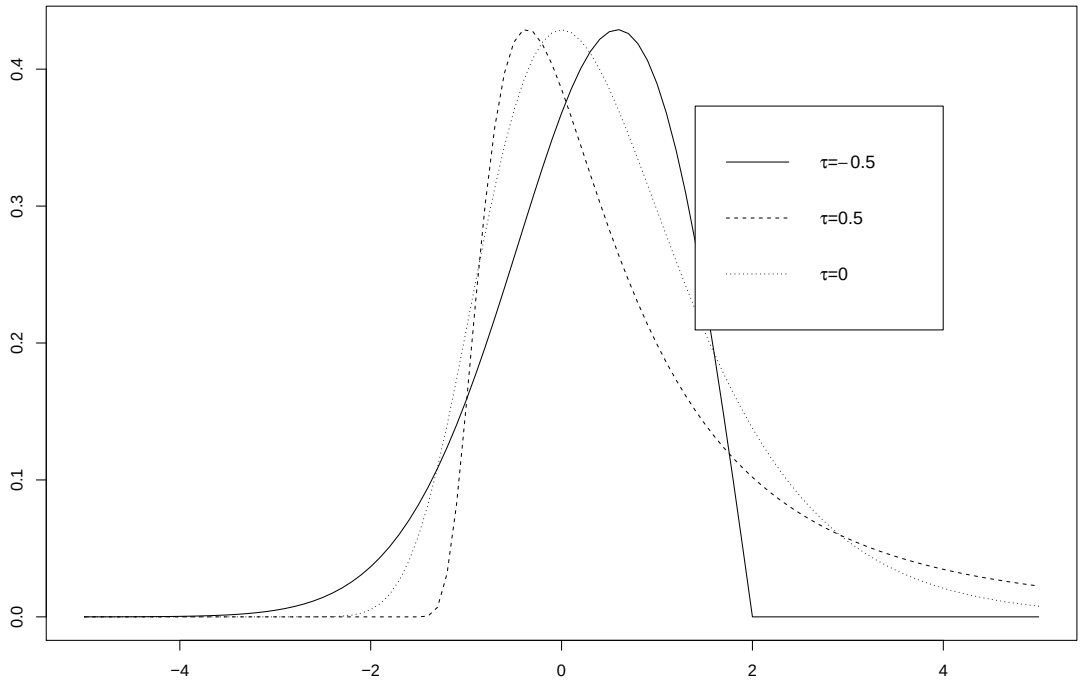

Figure 1: The GEV Density Functions for Different Values of the Shape Parameter τ.

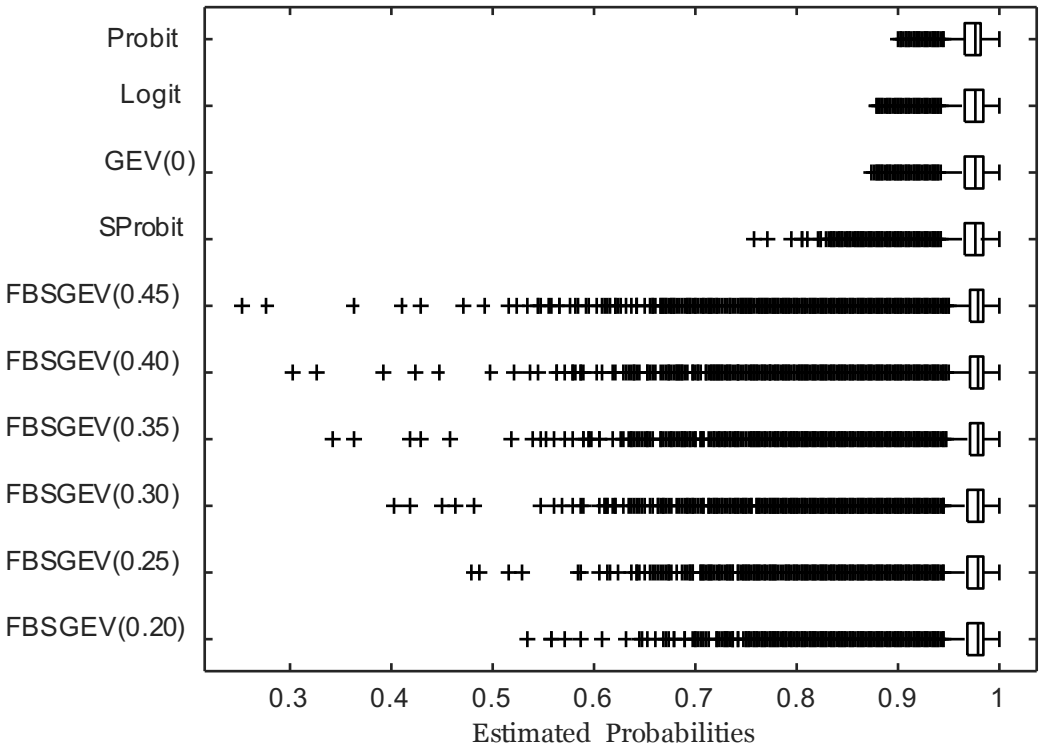

Figure 2: Box Plot of Estimated Probabilities of Repayment by Model based on the Data in 2009 (the default rate is 2.7%).

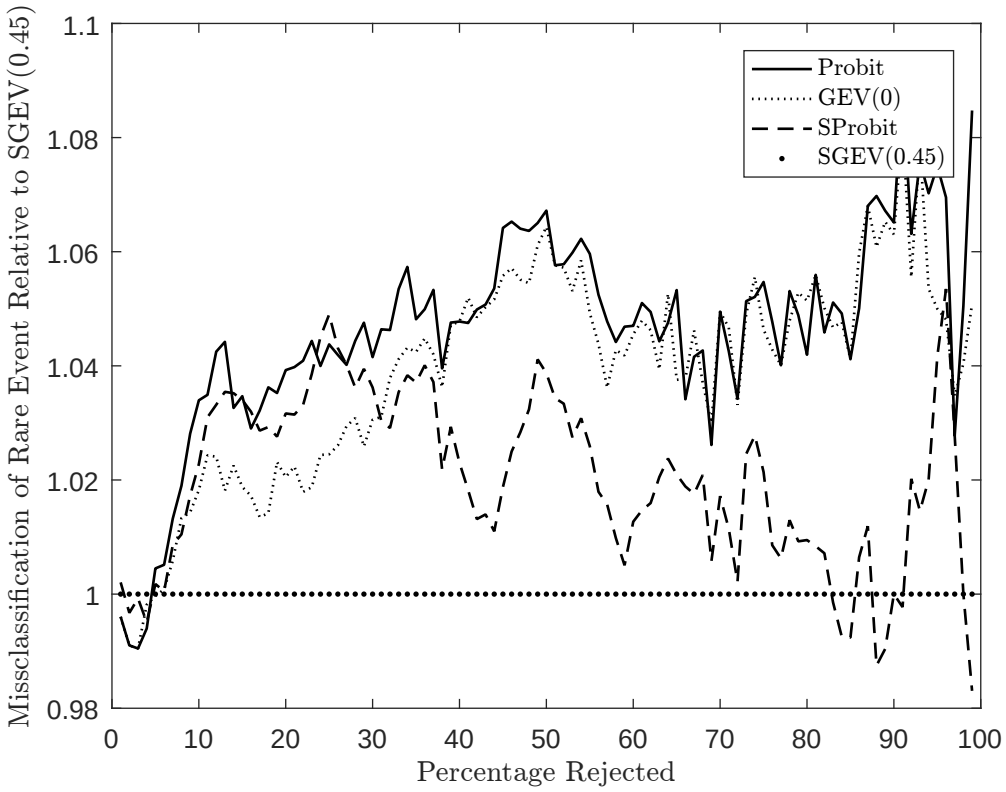

Figure 3: Relative Misclassification of Defaulted Mortgages by Model based on the Data in 2009 (the default rate is 2.7%).

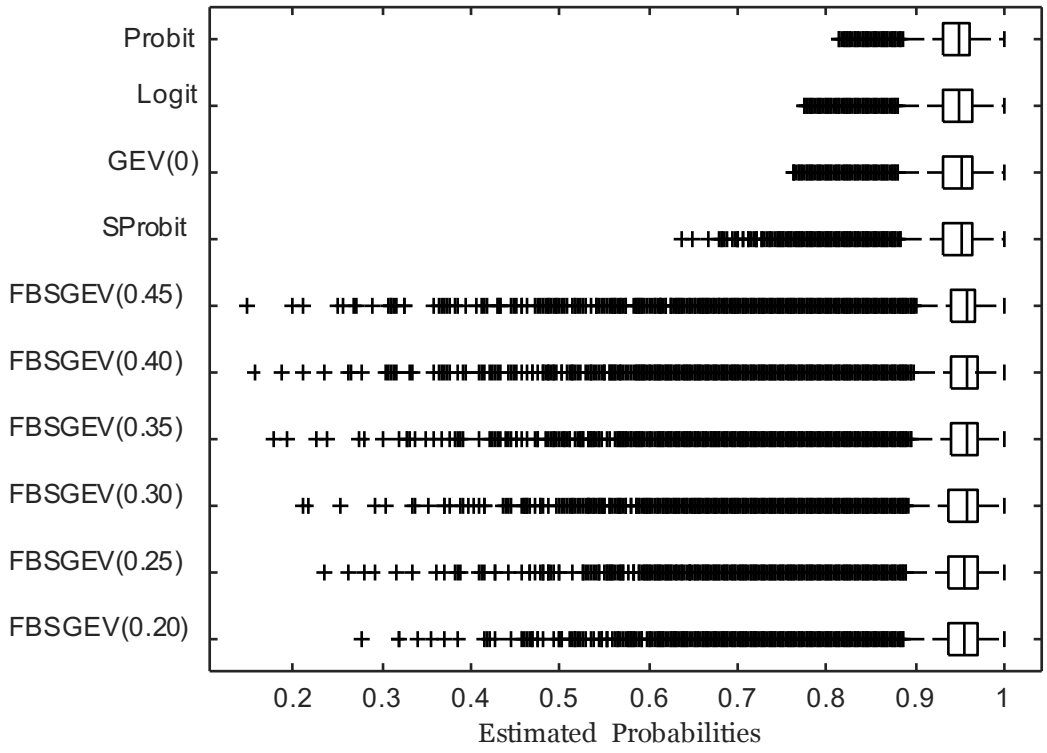

Figure 4: Box Plot of Estimated Probabilities of Repayment by Model based on the Data in 2009-10 (the default rate is 5.54%).

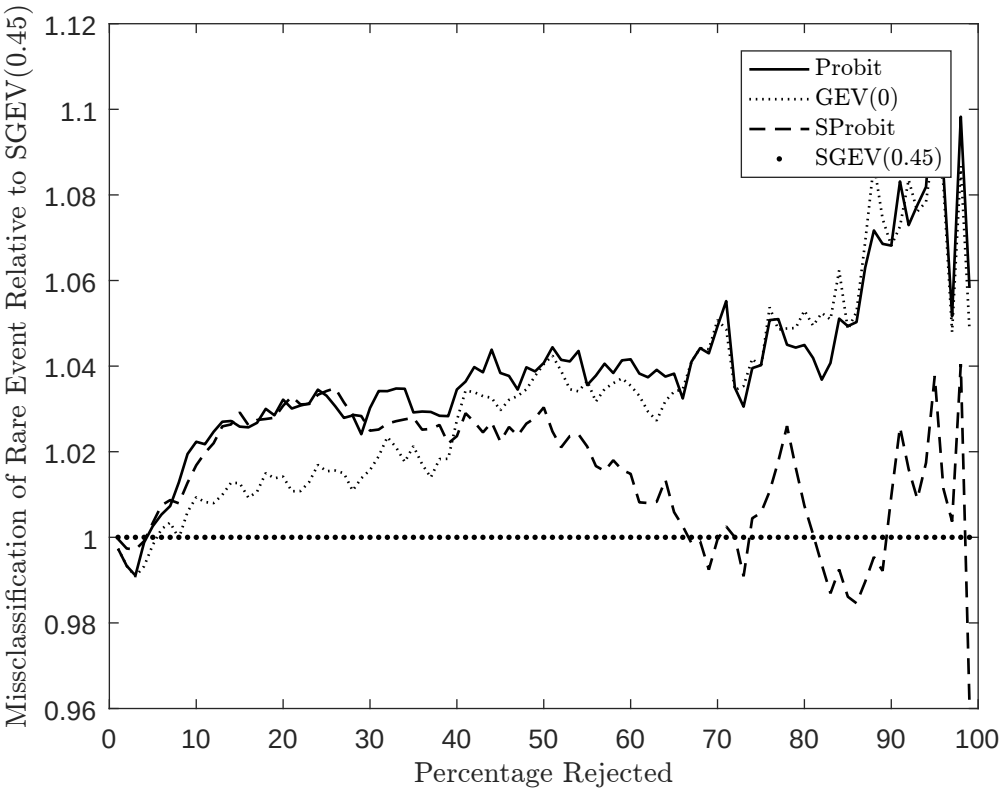

Figure 5: Relative Misclassification of Defaulted Mortgages by Model based on the Observations in 2009-10 (the default rate is 5.54%)