Spatial dependence in microfinance credit default

Spatial dependence in microfinance credit default

Victor Medina-Olivares[1], Raffaella Calabrese[1], Yizhe Dong[1], and Baofeng Shi[2]

[1]Business School, University of Edinburgh, 29 Buccleuch Place, Edinburgh EH8 9JS, victor.medina@ed.ac.uk, raffaella.calabrese@ed.ac.uk, yizhe.dong@ed.ac.uk

[2]College of Economics and Management, Northwest A&F University, Yangling, Shaanxi, 712100, China, shibaofeng@nwafu.edu.cn

Abstract

Credit scoring model development is very important for the lending decisions of financial institutions. The creditworthiness of borrowers is evaluated by assessing their hard and soft information. However, the microfinance borrowers are very sensitive to a local economic downturn and extreme (weather or climate) events. Therefore, this paper is devoted to extending the standard credit scoring models by taking into account the spatial dependence in credit risk. We estimate a credit scoring model with spatial random effects using the distance matrix based on the borrowers' locations. We find that including the spatial random effects improves the ability to predict defaults and non-defaults of both individual and group loans. Furthermore, we find that several loan characteristics and demographic information are important determinants of individual loan default but not group loans. Our study provides valuable insights for professionals and academics in credit scoring for microfinance and rural finance.

Keywords: Spatial dependence; credit scoring, microfinance, group lending.

1 Introduction

Microfinance, initially introduced in Bangladesh, is now widely considered as one of the most important innovations in development policy in the last four decades (Bayulgen, 2008;

Cull and Morduch, 2018). It aims to provide small loans to poor and low-income people who have limited or no access to the services provided by formal financial intermediaries to finance micro-businesses, build assets, and stabilise consumption. According to the 2017 Global Findex, about 1.7 billion adults still do not have access to formal financial services. Nearly all unbanked person live in developing economies. In China and India, for example, there are more than 225 million and 190 million of unbanked people, respectively. After years of rapid growth, various microfinance institutions (MFIs) are playing an important and positive role in developing human and social capital and improving the lives of poor people in the world (e.g. Abrar et al. (2021); Ganle et al. (2015); Imai et al. (2012); Sun and Liang (2021)). However, they also face many challenges. One of the major challenges faced is that they have been struggling with increased repayment problems among their borrowers. Microfinance borrowers are risky because they are typically low net-worth individuals who can offer little or no collateral for their loans. Therefore, understanding the factors affecting microfinance loan defaults is extremely important for the sustainable development of MFIs (Shi et al., 2016).

The main aim of this paper is twofold. First, we investigate whether spatial random effects in a scoring model for microfinance loans in China can improve the predictive accuracy of credit risk assessments. Second, our study identifies and compares the key drivers of credit risk in individual and group microfinance lending, which help MFIs make better credit underwriting decisions. To the best of our knowledge, this is the first study incorporating spatial effects into credit scoring models for individual and group microfinance loans.

To take into account that microfinance borrowers are affected by similar economic shock (McIntosh, 2008), we introduce spatial dependence in scoring models using the following techniques[1]. We first include only independent random effects in a logistic regression in line with Sohn and Kim (2007). We then consider spatial random effects using conditional autoregressive models introduced by Besag (1974). To measure the importance of the spatial dependence, we consider a mixture between a model with spatial random effects and one with independent random effects proposed by Leroux et al. (2000).We choose these models because they are widely used in the literature on spatial statistics (Banerjee et al., 2014). Given that the estimation of these models is computational intensive, we use the integrated nested Laplace approximations (INLA) (Rue et al., 2009).

We apply these approaches to two unique data sets on group and individual lending provided by a leading Chinese MFI. First, we obtain that individual and group lending show different risk drivers. We find that some variables related to loan size, repayment methods, marital status and loan purposes are significant to explain individual loan defaults. However,

[1] A growing number of studies provide evidence on the spatial dependence or credit contagion among medium-sized enterprises (SMEs) and mortgage borrowers (e.g. Calabrese et al. (2019); Babii et al. (2019); Fernandes and Artes (2016); **?**

the key drivers of the group loan default are completely different from those of individual loans. We find that loan size, borrowers from ethnic minorities, local unemployment rate and people that borrow more are important drivers for group loan defaults.Tables 14 and 15 show that the conditional default rate increases as the loan size increases for both individual and group loans. Second, adding independent random effects in a logistic regression model improves the calibration and the discrimination of scoring models for both individual and group lending. However, including spatial random effects outperforms a model with independent random effects only for individual loans. Finally, we estimate the model proposed by Leroux et al. (2000) and we obtain that the spatial component is significant for both the individual and the group lending.

The remainder of this paper is organised as follows. Section 3 is devoted to describing the methodology of spatial modelling for credit risk assessment. Section 4 describes the data used for the empirical analyses. Section 5 presents the empirical results and the effectiveness of the spatial models. Finally, Section 6 concludes this work.

2 Literature review

Assessing and managing credit risk is one of the most important tasks for financial institutions and regulators. The process of credit risk assessment is complex and unstructured. Many researchers and practitioners have developed a number of statistical and mathematical models to support lending decisions by transforming different types of data into numerical measures to discriminate "good" from "bad" loan applicants (Thomas, 2000; Thomas et al., 2017). Traditionally, financial and credit behavioural characteristics of borrowers are naturally considered as key risk factors of loan defaults (Altman and Sabato, 2007; Crone and Finlay, 2012; Wang et al., 2011). Many researchers find that demographic characteristics or other soft information are also related to the default probability of a credit applicant (Bravo et al., 2013; Cornée, 2019; Jiang et al., 2018).

Some studies also highlight the importance of incorporating macroeconomic conditions for the estimation of borrowers' credit risk (Bellotti and Crook, 2013). Carling et al. (2007) show macroeconomic variables have significant explanatory power for the default risk of firms. Bai et al. (2019) also find that the regional macroeconomic factors impact on farmers' credit risk. More recently, several studies find evidence of spatial dependence between loan defaults and suggest that **spatial contagion can improve the predictive accuracy of scoring models (Fernandes and Artes, 2016; Calabrese et al., 2019; Babii et al., 2019; Calabrese and Crook, 2020).**

Different models have been used to include spatial dependence in a scoring model for various loan products. Fernandes and Artes (2016) estimate a co-

variate that represents the spatial interdependence among borrowers using the kriging method that takes into account the distance among borrowers and the spatial interdependence associated to a given variable. They then include this variable in a logistic regression model for loans to small businesses. Barro and Basso (2010) instead use an entropy spatial interaction model which takes into account the distances between the regions where the firms are located and the economic sectors of these regions. Calabrese et al. (2019) use a completely different approach to incorporate spatial interactions in scoring models for small businesses given by the simultaneously autoregressive model (SAR) in a cross-sectional framework. Instead in a longitudinal context, Calabrese and Crook (2020) use a conditionally autoregressive model (CAR) in a survival approach for modelling mortgage defaults.

Although large efforts have been made to build credit scoring models, modelling credit risk for microfinance is more difficult than that for corporate and personal loans, largely because of insufficient or unverifiable information available (Dellien and Schreiner, 2005). Studies on credit scoring in the microfinance industry have not been sufficient. Bumacov et al. (2014, 2017) outline a conceptual framework of credit scoring for microfinance. Using data from a Bosnia–Herzegovinian microlender, Van Gool et al. (2012) develop logistic regression-based scoring models and suggest that credit scoring is a useful refinement tool in the microfinance lending process. Blanco et al. (2013) and Cubiles-De-La-Vega et al. (2013) implement several statistical credit scoring models based on the multilayer perceptron neural networks (MLP). Based on the data from the microfinance industry in Peru, they find neural network models outperform the other classic techniques such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and logistic regression (LR) techniques. Gicić and Subasi (2019) develop a microcredit scoring model based on synthetic minority oversampling technique (SMOTE) and ensemble classifier and they find that the proposed SMOTE model has a better prediction than other 57 different decision-support models including support vector machine (SVM), decision tree algorithms Bayesian networks. Serrano-Cinca et al. (2016) propose a financial and social decision-making model to evaluate the creditworthiness of microcredit borrowers in Colombia and suggest that the MFIs need to take social and environmental issues into account in their decision systems. Dorfleitner et al. (2017) examine drivers of the credit risk in Nicaraguan agricultural micro loans and find that the marital status, age and the economic objective of the loans significantly influence the probability of default.

Based on our knowledge, there is no paper in the literature that investigates the impact of spatial effects on a scoring model for microfinance. As the majority of the microfinace borrowers work in the agricultural or artisanal sectors, they are highly exposed to local economic shocks. For example, adverse weather events or environmental changes can significantly affect the crop, livestock productions and the borrowers' income, therefore increasing the default risk of the borrowers located in the same region (McIntosh, 2008).

3 Methodology

3.1 Spatial logistic models

Sohn and Kim (2007) propose a logistic regression model with random effects as a scoring model for small and medium enterprises. The main advantage of this approach lies in the ability of accommodating both the individual characteristics of each borrower and the uncertainty that is not explained by such individual factors. The authors also show that a model with random effects outperforms the same approach with fixed effects. In line with these findings, we introduce random effects in a logistic regression model.

Suppose Y_{ij} is a binary variable that takes the value of 1 if the borrower i ($i = 1, ..., n_j$) in the region j ($j = 1, ..., J$) defaults and 0 otherwise, and let $\mathbf{x}_{ij}$ denote the vector of borrower-specific covariates. Then we assume that the default event Y_{ij} follows

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ \text{logit}(p_{ij}) &= \boldsymbol{\beta}^{\intercal}\mathbf{x}_{ij} + W_j \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of coefficients associated with the covariates and W_j is the region-specific random effects. Denoting $\mathbf{y} = \{y_{ij}\}$, $\mathbf{x} = \{x_{ij}\}$ and $\mathbf{W} = \{W_j\}$, the likelihood of this model reads

$$\begin{aligned} L(\boldsymbol{\beta}, \mathbf{W}; \mathbf{y}, \mathbf{x}) &\propto \prod_{j=1}^{J}\prod_{i=1}^{n_i} p_{ij}^{y_{ij}}(1 - p_{ij})^{1-y_{ij}} \\ &= \prod_{j=1}^{J}\prod_{i=1}^{n_i} \{\text{logit}^{-1}(\boldsymbol{\beta}^{\intercal}\mathbf{x}_{ij} + W_j)\}^{y_{ij}}\{1 - \text{logit}^{-1}(\boldsymbol{\beta}^{\intercal}\mathbf{x}_{ij} + W_j)\}^{1-y_{ij}}. \end{aligned}$$

As we use a Bayesian approach, we consider the prior distributions $p(\mathbf{W}|\boldsymbol{\gamma})$, $p(\boldsymbol{\gamma})$ and $p(\boldsymbol{\beta})$ where $\boldsymbol{\gamma}$ is a hyperparameter for $\mathbf{W}$. We use a zero-mean Gaussian prior for the $\boldsymbol{\beta}$ coefficients with a precision equal to 0.001.

For the spatial random effects priors, we use three different configurations. The first one considers the random effects W_j as independent among regions j (Sohn and Kim, 2007), i.e. no spatial structure is considered, and distributed as a zero-mean Gaussian with precision τ, specifically

$$p(W_j|\tau) \sim \mathcal{N}\left(0, \tau^{-1}\right). \tag{2}$$

This distribution is then represented by only one hyperparameter $\gamma = \tau$. The prior for this hyperparameter is defined in its logarithmic scale and we use a log-gamma distribution with shape parameter of 1 and inverse-scale parameter of 0.0005 as a weak informative prior.

The second spatial prior we use, widely applied in spatial statistics (Banerjee et al., 2014), is a conditionally autoregressive (CAR) model (Besag, 1974) in its proper version.We choose a CAR model because it is computationally convenient to estimate complicated joint statistical relationships using a set of conditional dependencies (Banerjee et al., 2014). This model is defined as follows

$$p(W_j|W_{-j}, \tau_1, d) \sim \mathcal{N}\left(\frac{1}{d+m_j}\sum_{k\sim j} W_k, \frac{1}{\tau_1(d+m_j)}\right), \tag{3}$$

where W_{-j} represents the sets of spatial effects without region j, m_j is the number of neighbours of region j and $k \sim j$ indicates that the two regions k and j are neighbours. Equation 3 shows that the mean of the spatial random effect W_j is inversely proportion to m_j the number of neighbours of region j. This distribution is specified by two hyperparameters $\boldsymbol{\gamma} = (\tau_1, d)$. The term $\tau > 0$ is a precision-like parameter and d controls the "properness" of the covariance matrix where $d = 0$ corresponds to the improper CAR prior or also known as intrinsic CAR (ICAR) (Besag et al., 1991)[2]. The ICAR model is a common prior assumption in spatial lattice analysis, mainly because of computational advantages (Banerjee et al., 2014), but also is less flexible than this proper version. As before, the priors for these hyperparameters are defined in their logarithmic scale, both as a log-gamma distribution. For the τ_1 term we use the same as τ and for d term we use both parameters equal to 1.

Finally, the third spatial prior is the one proposed by Leroux et al. (2000). In this specification, the precision matrix is a convex combination of a CAR precision matrix Q and an identity matrix I which represents i.i.d. random effects. We choose this approach as it represents a weighted average of the first two specifications considered in this paper.

In analytical terms, we can write the precision matrix as $\tau_2(\lambda Q + (1-\lambda)I)$ where τ_2 is a precision parameter and $0 \leq \lambda \leq 1$ measures the strength of the spatial structure in the data. The conditional distribution for the random effects W_j follows

$$p(W_j|W_{-j}, \tau_2, \lambda) \sim \mathcal{N}\left(\frac{\lambda}{1-\lambda+\lambda m_j}\sum_{k\sim j} W_k, \frac{1}{\tau_2(1-\lambda+\lambda m_j)}\right). \tag{4}$$

Large values of λ indicate a strong spatial pattern(for the limit value $\lambda = 1$, we obtain a CAR model), while small values of λ show a weak spatial pattern (for the limit value $\lambda = 0$, we obtain independent random effects). The prior of τ_2 is analogous to the ones defined above and for λ the prior is defined in its logit scale by a Gaussian distribution centred in zero with a precision of 0.45 (equivalent to a standard deviation of 1.5 so that the range of the prior of λ is constrained to the interval (0,1)).

Originally, we implemented the aforementioned models in the platform for statistical modelling *Stan*. Even though we used the No-U-Turn Sampler (Hoffman and Gelman,

[2]The improper CAR prior does not implied that the posterior is also improper.

2014), a faster extension to Hamiltonian Monte Carlo algorithm (HMC), we encountered high computational costs. We implemented the models using the integrated nested Laplace approximations (INLA) (Rue et al., 2009), a deterministic algorithm that provides accurate and fast Bayesian inference and is described in the following section.

3.2 INLA methodology

The novelty of INLA methodology can be summarised in two main characteristics. First, it focuses on estimating the posterior marginal distributions of the parameters, rather than the joint posterior which is difficult to obtain, specially if we deal with a high dimensional space. Second, it is suitable for models that can be expressed as latent Gaussian Markov random fields (GMRF) since this provides computational advantages in the inference (see Rue and Held (2005)).

In our case, we define the vector of latent effects $\mathcal{X} = (\{\eta_{ij}\}, \boldsymbol{\beta}, \mathbf{W})$, with $\eta_{ij} = \boldsymbol{\beta}^\intercal \mathbf{x}_{ij} + W_j$. Then, assuming that $\mathcal{X}$ is a GMRF and noting that the observations $\mathbf{y}$ are independent given $\mathcal{X}$, we can estimate this model with INLA. The distributions of the elements of $\mathcal{X}$ depend on a vector of hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \theta_{\boldsymbol{\beta}})$ where $\theta_{\boldsymbol{\beta}}$ denotes the hyperparameters for $\boldsymbol{\beta}$. Formally, we have

$$\mathbf{y}|\mathcal{X}, \boldsymbol{\theta} \sim \prod_{k \in \mathcal{K}} p(y_k|\mathcal{X}_k, \boldsymbol{\theta}),$$

where $\mathcal{K}$ is the set of indices for all observed values in $\mathbf{y}$ and it is coded so that each observation is associated with its respective element $\mathcal{X}_k$. We assume the density of $\mathcal{X}|\boldsymbol{\theta}$ as zero-mean Gaussian with precision matrix $\mathbf{Q}(\boldsymbol{\theta})$. Therefore, the posterior distribution follows

$$\begin{aligned} p(\mathcal{X}, \boldsymbol{\theta}|\mathbf{y}) &\propto p(\boldsymbol{\theta}) p(\mathcal{X}|\boldsymbol{\theta}) \prod_{k \in \mathcal{K}} p(y_k|\mathcal{X}_k, \boldsymbol{\theta}) \\ &\propto p(\boldsymbol{\theta}) |\mathbf{Q}(\boldsymbol{\theta})|^{1/2} \exp\left[-\frac{1}{2}\mathcal{X}^\intercal \mathbf{Q}(\boldsymbol{\theta})\mathcal{X} + \sum_{k \in \mathcal{K}} \log\{p(y_k|\mathcal{X}_k, \boldsymbol{\theta})\}\right]. \end{aligned}$$

The posterior marginal distributions, $p(\mathcal{X}_k|\mathbf{y})$ and $p(\theta_j|\mathbf{y})$, are given by

$$\begin{aligned} p(\mathcal{X}_k|\mathbf{y}) &= \int p(\mathcal{X}_k|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \\ p(\theta_j|\mathbf{y}) &= \int p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_{-j}. \end{aligned} \tag{5}$$

INLA methodology approximates these posterior marginals by using the Laplace method (Tierney and Kadane, 1986). For the term $p(\boldsymbol{\theta}|\mathbf{y})$ the approximation corresponds to

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}) &\propto \frac{p(\mathcal{X},\boldsymbol{\theta},\mathbf{y})}{p(\mathcal{X}|\boldsymbol{\theta},\mathbf{y})} \\ &\approx \left.\frac{p(\mathcal{X},\boldsymbol{\theta},\mathbf{y})}{\tilde{p}_G(\mathcal{X}|\boldsymbol{\theta},\mathbf{y})}\right|_{\mathcal{X}=\mathcal{X}^*(\boldsymbol{\theta})} =: \tilde{p}(\boldsymbol{\theta}|\mathbf{y}) \end{aligned}$$

where $\tilde{p}_G(\mathcal{X}|\boldsymbol{\theta},\mathbf{y})$ denotes the Gaussian approximation to the full conditional and $\mathcal{X}^*(\boldsymbol{\theta})$ its mode. A further Laplace approximation is done for the terms $p(\mathcal{X}_k|\boldsymbol{\theta},\mathbf{y})$ as follows

$$\begin{aligned} p(\mathcal{X}_k|\boldsymbol{\theta},\mathbf{y}) &\propto \frac{p(\mathcal{X},\boldsymbol{\theta},\mathbf{y})}{p(\mathcal{X}_{-k}|\mathcal{X}_k,\boldsymbol{\theta},\mathbf{y})} \\ &\approx \left.\frac{p(\mathcal{X},\boldsymbol{\theta},\mathbf{y})}{\tilde{p}_G(\mathcal{X}_{-k}|\mathcal{X}_k,\boldsymbol{\theta},\mathbf{y})}\right|_{\mathcal{X}_{-k}=\mathcal{X}^*_{-k}(\mathcal{X}_k,\boldsymbol{\theta})} =: \tilde{p}(\mathcal{X}_k|\boldsymbol{\theta},\mathbf{y}). \end{aligned}$$

Ultimately, the terms in Equation 5 are replaced by their corresponding approximations and the integrals are computed using numerical methods.

4 Data set

The data employed in our empirical study is rather unique and provided by one of the largest Chinese microfinance institutions (MFIs). The MFI was formally established in 2008 and offers micro-finance services to poverty-stricken populations in rural areas for production and consumption. By the end of 2019, it had 344 business outlets in 21 provinces across China and has cumulatively granted more than CNY 57 billion (USD 8.16 billion) loans, with an average loan size of CNY 26,000 **(USD 3,720)**. The data set contains two types of loans, namely individual and group loans, during January 2017 to July 2018. It covers 240 counties in 20 Chinese provinces. Other characteristics are illustrated as follows:

- Individual loans: This sample consists of 8,513 loans granted to individuals with an average loan value of CNY 45,333 (USD 6,488)and a mean term of 12 months. The default rate of this data set of loans is 3.2%. A default is defined as the borrower being 30 or more days delinquent. Top three purposes of loans are wholesale and retail trades, social service, and crop farming. They account for 16.9%, 16.4% and 16.0% of total loans, respectively.
- Group loans: This sample includes 15,348 group loans. Under joint liability, small groups of borrowers are responsible for the repayment of each other's loans. Group members are treated as being in default if at least one of them does not repay and all members are then denied subsequent. This means that the default label is provided

at the group level, i.e. we cannot differentiate between group members. We follow the MFI's default criterion and define a loan in default if the loan payment is overdue for 30 or more days. The group averagely consists of 3 members. The averages of loan value and term are, respectively, CNY 15,978 (USD 2,286) and 10 months. Crop farming, raising livestock and breeding fish, and wholesale and retail trades are the three main purposes of borrowing, accounting in total for 64.9% of all loans. The data set presents a default rate of 2.4%

Figure 1 shows the spatial distribution of the individual and group borrowers analysed in this paper. We note that North China, Northeast China and Northwest China have a greater density of group borrowers while individual borrowers are more clustered in Central China, Southwest China and Beijing area.

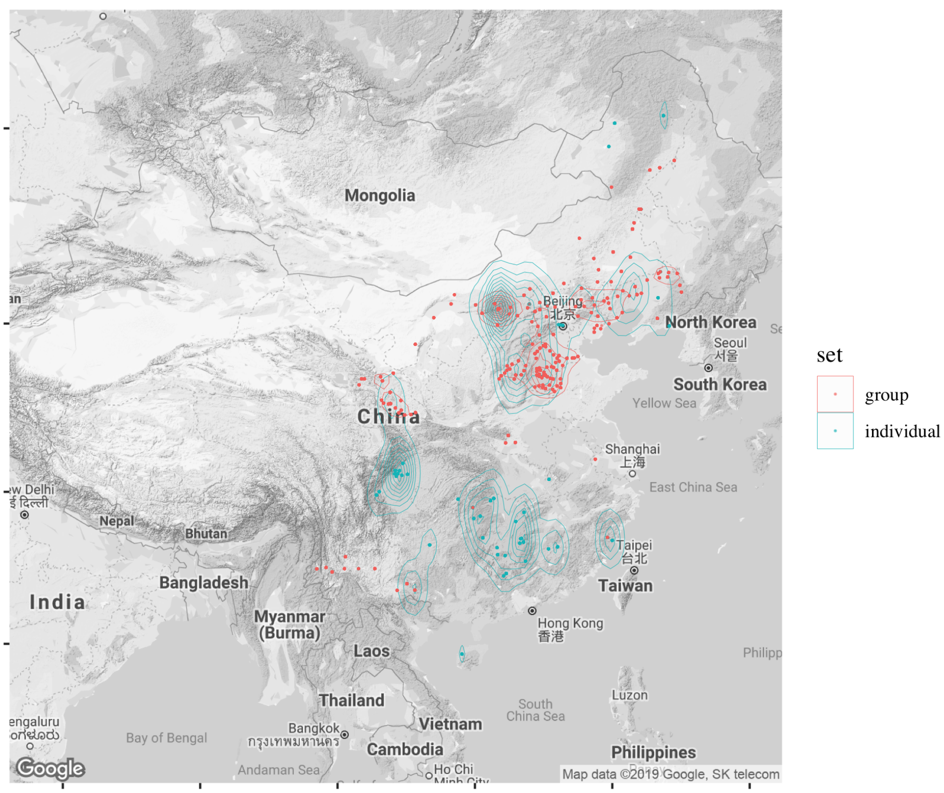

Figure 1: Spatial distribution for the observations with kernel density estimates.

Both data sets have 20 variables in total, 11 numeric and 9 categorical, in addition to the location of the borrowers. Among these variables we have the loan purpose, the amount and its term, repayment type, borrower demographics, among others, which are in line with the literature (Li and Zhenyu, 2018). Coherently with (Bellotti and Crook, 2009), we added annual local macroeconomic indicators at city level.Local macroeconomic variables are collected from Data China provincial statistical yearbooks. The detailed definition of each variable is reported in Table 1.

5 Empirical results

For prediction purposes, we randomly select 75% of the borrowers as training set where we obtain the results for Sections 5.1 and 5.2 and 25% as test set that we use to compute the prediction performance metrics shown in Section 5.3. We select the training and test sets using a stratified random sampling where the strata are based on the default status. We report the descriptive statistics for the training and test sets in Tables 17 and 18 in the Appendix C.

5.1 Determinants of credit default risk for individual and group loans

The identification of credit risk key drivers is crucial in lending decision-making processes for MFIs and policymakers. Therefore, we are investigating what are the main factors of microfinance defaults. Tables 2 and 3 present the estimation results of the estimated models after the exclusion of non-significant variables (above 10% significant level[3]) for the individual and group lending samples, respectively[4].

For individual loans, we find that several loan characteristics and borrower demographic information have a significant impact on microfinance loan default. Specifically, loans with an annuity repayment structure (i.e. repayment 1 and 2) are less likely to default than monthly payment structure (repayment 0), suggesting the MFIs could reduce individual loan defaults by redesigning their loan repayment methods. Meanwhile, we find that larger loan size, borrowers belonging to ethnic minority and loans used for consumption (purpose 7) are more likely to default. As the main aim of microfinance is to target non-bankable borrowers who usually do not have access to mainstream financial markets (Pollinger et al., 2007), we include gender and ethnic minority as explanatory variables. This choice would also allow us to explore sources of lack of fairness and discrimination in this credit system (Kamishima et al., 2012; Cheng, 2015).

We evidence that loans used for wholesale and retail trades (purpose 3) are also more likely to defaults but with not such a strong significance. Moreover, we find that the local amount of loans per capita and deposits from financial institutions, both indicators of the

[3]Even if we use a Bayesian approach to estimate the parameters of the models, we apply a frequentist method for the variable selection. We estimate a standard logistic and we choose the explanatory variables with 10% significance level.

[4]While our approach to include random effects is Bayesian, the procedure we used to check if the independent variables are multicollinear is frequentist. Particularly, we compute the Variance Inflation Factor (Fox and Monette, 1992) and we obtain for all the variables a value around one. This means that there is no multicollinearity between the covariates.

use of banking services, are important predictors of individual loan default.

The results of Table 3 show that there are less risk factors affecting group lending defaults compared with individual lending. Specifically, we find loan size, ethnic minority borrowers, local unemployment rate and the education level are positively related to group loan default. The results also indicate that, coherently with the expectations, the group lending mechanism (e.g. joint liability, group reputation, and future access to credit for each member) makes the loan and borrowers' characteristics less important to determine the loan default.

5.2 Spatial random effects in credit scoring

In line with Sohn and Kim (2007), we add random effects to a logistic regression model. First, we consider independent random effects described in Equation 2. The results of this analysis are shown in Table 4 for individual loans, and in Table 5 for group loans. We compare the results obtained from the models without random effects in Tables 2 and 3 with those with independent random effects in Tables 4 and 5. We notice that the signs of the posterior mean of the parameter estimates in the models with and without random effects coincide but the value could change, in some cases substantially. For some variables, the credible intervals for the model without and with random effects are similar, for example for a divorced borrower (i.e. $marital2$) the two credible intervals are similar (0.529, 1.370) and (0.511, 1.353). Instead these intervals can be different for other variables, for example for $ethnicother$ the two intervals are respectively (0.220, 1.522) and (0.157, 1.477). For the group loans, we notice that the $purpose\_loan10$ and $ethnicother$ have a significant effect when no random effects are considered, but not anymore when we include the random effects. The opposite happens to $ethnic5$, which is slightly relevant in the case with no random effects but, with random effects, its effect has almost doubled. Some of these variables can incorporate spatial effects, so their estimates can change when random effects based on the geographical areas are included in the model.

Afterwards, we add spatial random effects with the two different prior assumptions, CAR and Leroux, described in Equations 3 and 4, respectively. In spatial statistics there are different approaches to define a neighbourhood. In this analysis, we know the longitude and latitude of the centroid of the county where each borrower is resident. Since counties differ considerably in size, we define two borrowers as neighbours if the distance between the two centroids is less than or equal to 70 km. We chose this value because it is approximately the median of all the distances between each pair of centroids in the data set (see Appendix B for details). The results for the individual loans with the CAR random effects are in Table 6 and with the Leroux random effects in Table 7. The parameter estimates for both models show consistent sign and magnitude, however the hyperparameters tell us additional information about the spatial structure. In the case of the CAR specification, for example, we observe that d departs from zero. In other words, the representation of the spatial structure is not well explained by the widely used ICAR model. With respect to the Leroux specification, we see in Table 7 that the parameter λ, which measures the evidence of the spatial structure, is significant for this data set (mean value of 0.351). Figure 2 show maps of the posterior mean of the spatial random effects for the independent, CAR and Leroux random effects models for individual data set As the patterns from the three plots suggest that there are similarity in spatial effects among three models.

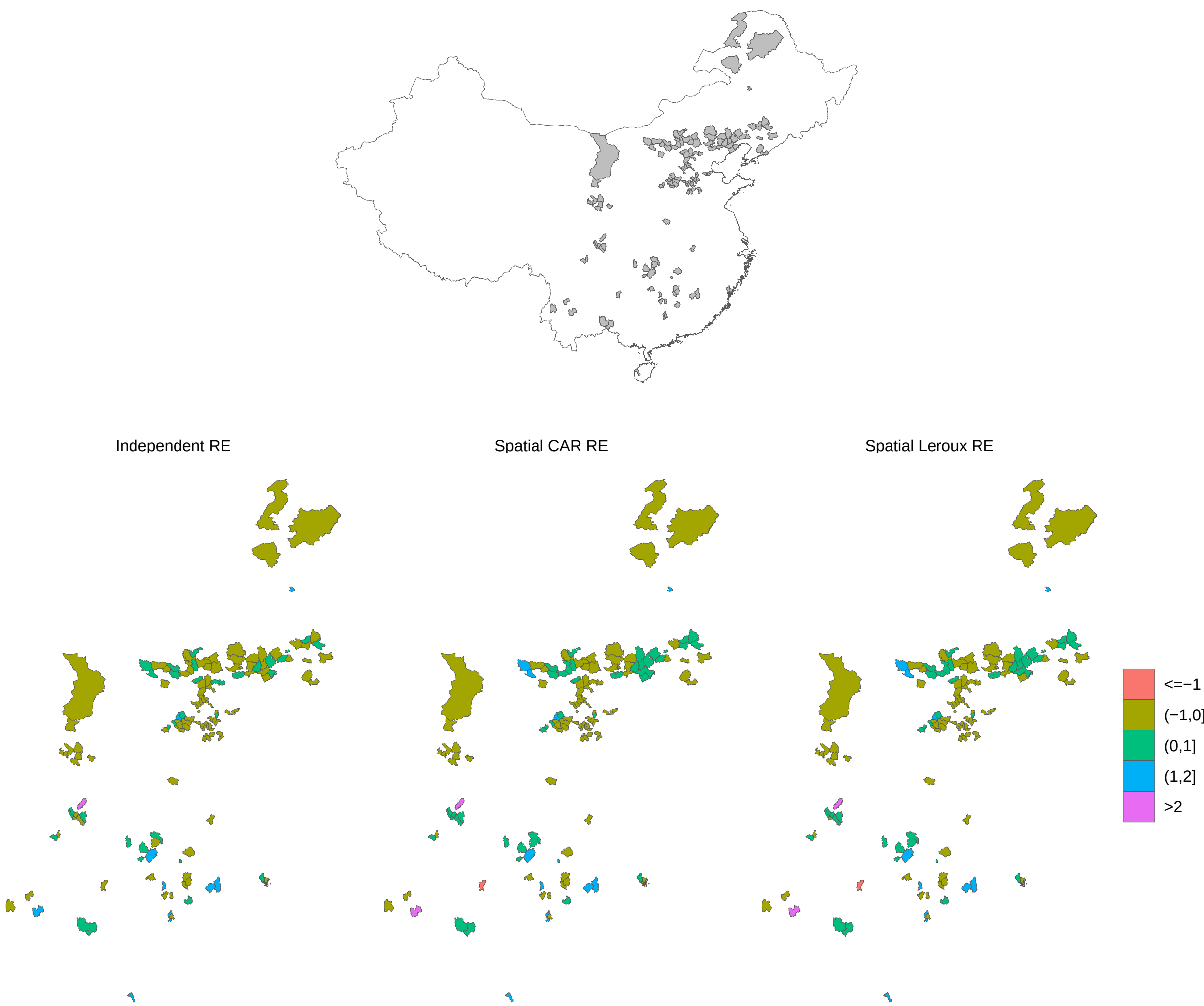

Figure 2: The posterior mean of the spatial random effects for individual loans.

The results for the group loans are shown in Tables 8 and 9 for CAR and Leroux random effects, respectively. We observe, as in the case of individual loans, that when we include either the CAR or Leroux random effects, the parameters estimates do not change significantly. Moreover, we also observe that the spatial structure is better represented by a proper CAR model rather than improper (posterior mean of d equals to 2.154). Regarding the value of λ (Table 9), we also find evidence of spatial structure in contrast to the independence assumption. If we compare these results with those obtained in Tables 4 and 5 with independent random effects, we see that the mean of the estimates are similar. Figure 3 display maps of the posterior mean of the spatial random effects for the independent, CAR and Leroux random effects models based on group data set. The three plots show a very similar patterns to that of spatial random effects from different models.

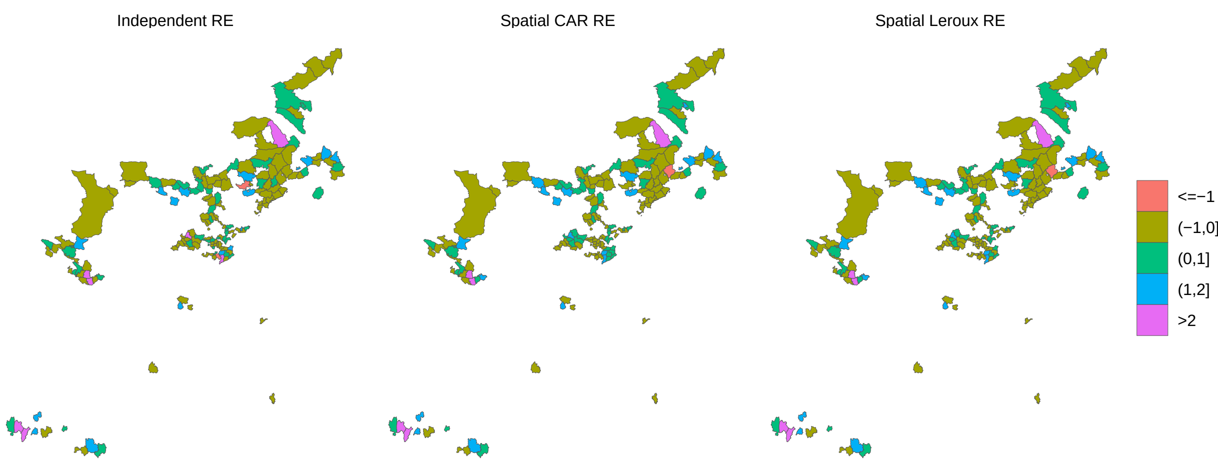

Figure 3: The posterior mean of the spatial random effects for group loans.

To measure the goodness of fit, we compute the Deviance information criterion (DIC) (Spiegelhalter et al., 2002), Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010) and the conditional predictive ordinate (CPO) (Geisser, 1980), widely used criteria for Bayesian model assessment. Lower values of DIC, WAIC and CPO imply better fit.

For the individual loans, we can see in Table 10 that the model with no random effects show the highest DIC, WAIC and CPO and therefore the worst fit to the data. Further, we observe that among the three models with random effects, the CAR and the Leroux specifications show similar values of DIC, WAIC and CPO and can represent the data better than the model with i.i.d. random effects.

On the contrary, Table 11 shows that the best model for group loans is the one with independent random effects. This result might be due to ignoring the different locations of the borrowers in the group lending but considering only the location of the main borrower in the group.

5.3 Performance metrics in out-of-sample analysis

In previous section, we saw that the random spatial effects improve the in-sample accuracy only for individual loans. We are now interested in exploring how the spatial effect influences the prediction performance (out-of-sample).

Using the credit scoring models presented in Section 5.1, we estimate the probabilities of default in the out-of-sample data sets with and without spatial effects. We use the Bayesian predictive distribution to estimate the predictions assuming that both the observed and the new realisations are exchangeable(see Ch.5 Blangiardo and Cameletti (2015) for further details). We analyse the models using discrimination and calibration measures. Discrimination indices, such as the Area Under the ROC Curve (AUC) (Fawcett, 2006; Zou et al., 2007), the H index (Hand, 2009), the Gini, and the Kolmogorov-Smirnoff (KS) statistic, try to measure how well the model can distinguish between defaulted and non-defaulted borrowers. Instead, a calibration index, such as the logarithmic score (Good, 1952), measures how close are the estimated probabilities to the true values. In Table 12 we obtain that the H index for the model with spatial random effects shows a slight improvement compared to the same index for the model without random effects and with independent random effects on individual loans. The AUC, Gini and the KS statistic show instead that the model with independent random effects outperforms the others. The main advantage of the H index on the AUC, Gini and the KS statistic is to take into account the imbalance of the binary dependent variable. The logarithmic score shows that the models with the spatial random effects provide closer probabilities to the true values.

The results for the group loans are shown in Table 13. In this case, the H index, the Gini, the AUC and the logarithmic score agree that the best performance is shown by the models with spatial random effects.

6 Conclusions

Microfinance has attracted significant interest in recent years, both from policy makers, industry practitioners as well as academics, as it plays a vital role in the poverty reduction of developing countries. It provides an opportunity for rural and poor people to access financial services. However, many MFIs are faced with the problem of high loan default rate because of the high risk profile of the individuals and micro-business they lend money to. Building an efficient and effective credit scoring model that is tailored to the particular characteristics of microfinance is extremely important. Traditionally, the creditworthiness of borrowers is assessed based on demographic characteristics, financial and repayment information. However, we recognise that these factors may not be enough to explain all the risks associated with microcredit borrowers. Therefore, in this study, we introduce a credit scoring model with spatial random effects.

We analyse the data provided by a leading Chinese MFI and we obtain that group and individual lending show different risk drivers. We find that some variables related to loan size, repayment methods, marital status and loan purposes are significant to explain individual loan defaults. However, the key drivers of the group loan default are completely different from those of individual loans. We find that loan size, borrowers from ethnic minorities and local unemployment rate are important drivers for group loan defaults. The fewer determinants of group loan default indicate that group lending may play a positive role in mitigating the risks associated with information asymmetry through joint liability.

In line with Sohn and Kim (2007), we then add independent random effects in a logistic regression model and we obtain a better goodness of fit than the model without random effects and, in terms of prediction, we also see improvements in both the calibration and discrimination metrics. We also consider spatial random effects under a CAR model specification but this approach outperforms the out-of-sample performance of the model with independent random effects only for group loans. Afterwards, we estimate a mixture between a model with spatial random effects and one with independent random effects proposed by Leroux et al. (2000). We obtain that the spatial component is significant for both the individual and the group lending. As all the models with random effects are computational intensive, we use the integrated nested Laplace approximations (INLA) (Rue et al., 2009).

This study has potential policy implications. In showing the importance of the spatial dependence for predicting credit defaults, our work suggests that policymakers need to work for stability in the macro-environment to help MFIs to increase social sustainability by providing more services to particular clients, while maintaining the financial and operational sustainability of the institutions. We also suggest that according to local conditions, MFIs could develop diversified credit products to meet the credit demands of different groups of people, such as borrowers from ethnic minorities.

A Default rates by loan size

Tables 14 and 15 show the default rate per loan size for four groups with equal range.

B Distribution of minimum distances

Figure 4 shows the boxplot of the distribution of minimum distances between all the pairs of centroids.

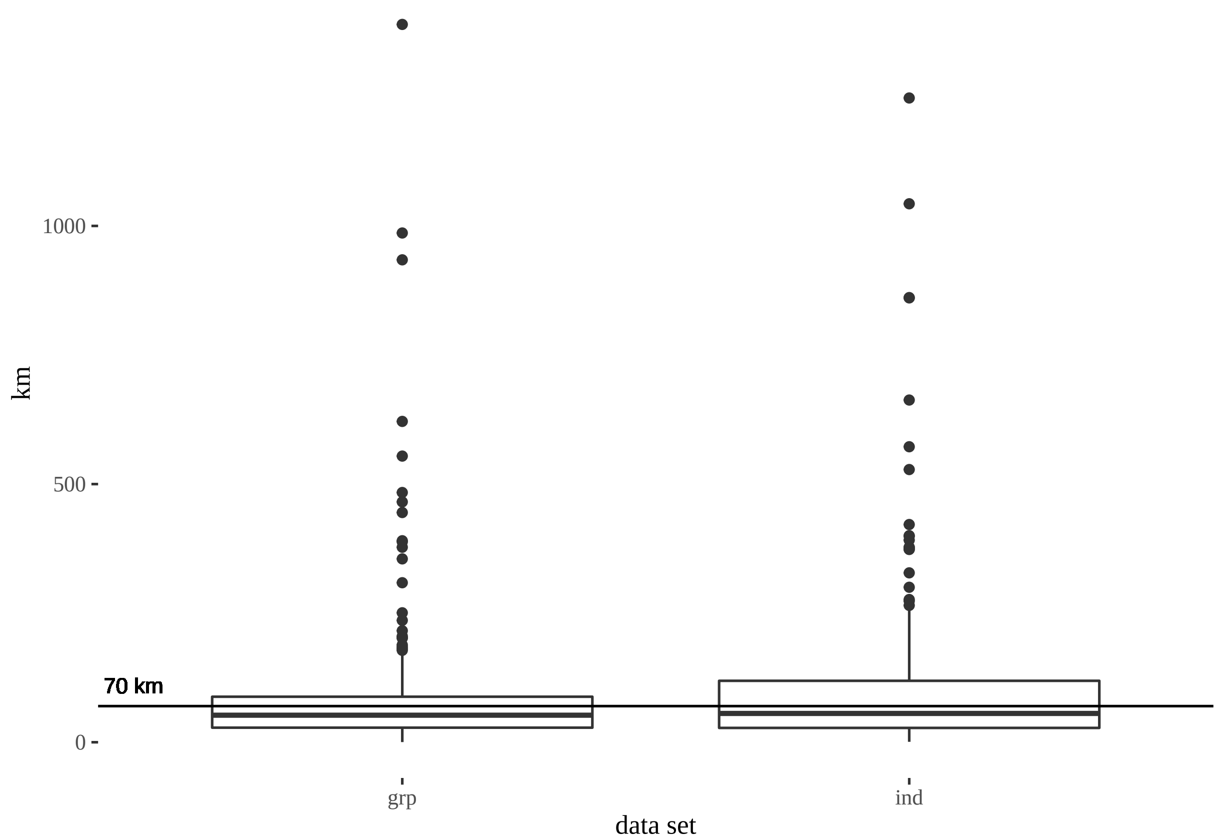

Figure 4: Boxplot of the distribution of minimum distances for each data set. The horizontal line indicates the ratio used to define the neighbours.

Using this ratio, the distribution of the number of links is shown in Table 16. From the table we see that for both data sets there are 29 locations without any neighbour. Also, the most connected location for the individual data set has 11 neighbours and 16 for the group data set.

C Descriptive statistics for the training and test sets.

Tables 17 and 18 show some descriptive statistics on training (75%) and test sets (25%).

D Codes

```
## Load packages
library(spdep)
library(rgdal)
library(tidyverse)
library(readxl)
library(INLA)

## Load data set
nam_type <- read_excel("./data/name_type.xlsx")
data <- read_excel("./data/data_ind.xlsx",
                   col_types = nam_type$type,
                   col_names = nam_type$name,
                   skip = 1)
## Distance matrix
coords <- data %>%
  count(long, lat, sort = T) %>%
  mutate(n_area = row_number())
sub_coords <- sp::coordinates(coords[c("long","lat")])
ratio <- 70
# identifies neighbours within ratio
nb_neigh <- dnearneigh(sub_coords, d1 = 0, d2 = ratio, longlat = TRUE)

# Build weight matrix
W <- nb2mat(nb_neigh, style="B", zero.policy = TRUE)

## add areas number
data <- data %>%
  left_join(coords, by = c("long", "lat"))
```

```
## Create out of sample
set.seed(1907)
id_test <- data %>%
  select(id, y) %>%
  group_by(y) %>%
  nest() %>%
  mutate(sample = map(data, sample_frac, .25)) %>%
  select(-data) %>%
  unnest(sample) %>%
  arrange(id) %>%
  pull(id)

data <- data %>%
  mutate(y_m = ifelse(id %in% id_test, NA, y))

## Define base model
ff <- as.formula(paste("y_m ~",
                       paste(names(data)[1:(ncol(data)-4)],
                             collapse = "+")))

## Model with no random effects
mod_simple <- inla(ff,
                   family ='binomial',
                   control.family = list(link = "logit"),
                   verbose = TRUE,
                   data = data,
                   control.compute = list(dic = TRUE,
                                          waic = TRUE,
                                          cpo = TRUE),
                   control.predictor = list(compute = TRUE,
                                            link = 1))
summary(mod_simple)
data$y_simple <- mod_simple$summary.fitted.values$mean

## Model area iid
mod_area_iid <- inla(update(ff, .~.  +  f(n_area, model = "iid")),
                     family = 'binomial',
                     control.family= list(link = "logit"),
                     verbose = TRUE,
```

```
                     data = data,
                     control.compute = list(dic = TRUE,
                                            waic = TRUE,
                                            cpo = TRUE),
                     control.predictor = list(compute = TRUE,
                                              link = 1))

summary(mod_area_iid)
data$y_ar_iid <- mod_area_iid$summary.fitted.values$mean

## Model proper CAR
mod_spat <- inla(update(ff, .~. + f(n_area, model = "besagproper",
                                    graph = W)),
                 family = 'binomial',
                 control.family= list(link = "logit"),
                 verbose = TRUE,
                 data = data,
                 control.compute = list(dic = TRUE,
                                        waic = TRUE,
                                        cpo = TRUE),
                 control.predictor = list(compute = TRUE,
                                          link = 1))
summary(mod_spat)
data$y_spat <- mod_spat$summary.fitted.values$mean

## Leroux Model
mod_spat_v2 <- inla(update(ff, .~. + f(n_area, model = "besagproper2",
                                       graph = W)),
                    family = 'binomial',
                    control.family= list(link = "logit"),
                    verbose = TRUE,
                    data = data,
                    control.compute = list(dic = TRUE,
                                           waic = TRUE,
                                           cpo = TRUE),
                    control.predictor = list(compute = TRUE,
                                             link = 1))
summary(mod_spat_v2)
data$y_spat_v2 <- mod_spat_v2$summary.fitted.values$mean
```

```
## Compare models
model_assess <- t(cbind(c(mod_simple$dic$dic,
                          mod_simple$waic$waic,
                          -sum(log(mod_simple$cpo$cpo), na.rm = T)),
                        c(mod_area_iid$dic$dic,
                          mod_area_iid$waic$waic,
                          -sum(log(mod_area_iid$cpo$cpo), na.rm = T)),
                        c(mod_spat$dic$dic,
                          mod_spat$waic$waic,
                          -sum(log(mod_spat$cpo$cpo), na.rm = T)),
                        c(mod_spat_v2$dic$dic,
                          mod_spat_v2$waic$waic,
                          -sum(log(mod_spat_v2$cpo$cpo), na.rm = T)) ))

rownames(model_assess) <- c("No RE",
                            "Independent RE",
                            "Spatial CAR RE",
                            "Spatial Leroux RE")
colnames(model_assess) <- c("DIC","WAIC","CPO")


## Discrimination and calibrarion metrics (25% sample)
labels <- data %>%
  filter(is.na(y_m)) %>%
  pull(y)
scores <- data %>%
  filter(is.na(y_m)) %>%
  select(y_simple, y_ar_iid, y_spat, y_spat_v2) %>%
  as.matrix()
log_score <- data %>%
  filter(is.na(y_m)) %>%
  select(y, y_simple, y_ar_iid, y_spat, y_spat_v2) %>%
  pivot_longer(-y) %>%
  mutate(value2 = -y*log(value)-(1-y)*log(1-value)) %>%
  group_by(name) %>%
  summarise(logs = mean(value2)) %>%
  slice(c(2,1,3,4))

met_table <- cbind(hmeasure::HMeasure(labels, scores)$metrics[,c(1,2,3,5)],
                   log_score[,2])
```

```
rownames(met_table) <- c("No RE",
                         "Independent RE",
                         "Spatial CAR RE",
                         "Spatial Leroux RE")
colnames(met_table) <- c("H-index","Gini","AUC","KS","Log Score")
```

References

Abrar, A., Hasan, I., and Kabir, R. (2021). Finance-growth nexus and banking efficiency: The impact of microfinance institutions. *Journal of Economics and Business*, 114:105975.

Altman, E. I. and Sabato, G. (2007). Modelling credit risk for smes: Evidence from the us market. *Abacus*, 43(3):332–357.

Babii, A., Chen, X., and Ghysels, E. (2019). Commercial and residential mortgage defaults: Spatial dependence with frailty. *Journal of econometrics*, 212(1):47–77.

Bai, C., Shi, B., Liu, F., and Sarkis, J. (2019). Banking credit worthiness: Evaluating the complex relationships. *Omega*, 83:26–38.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

Barro, D. and Basso, A. (2010). Credit contagion in a network of firms with spatial interaction. *European Journal of Operational Research*, 205(2):459—-468.

Bayulgen, O. (2008). Muhammad Yunus, Grameen Bank and the Nobel Peace Prize: What political science can contribute to and learn from the study of microcredit. *International Studies Review*, 10(3):525–547.

Bellotti, T. and Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, 60(1):1699–1707.

Bellotti, T. and Crook, J. (2013). Forecasting and stress testing credit card default using dynamic models. *International Journal of Forecasting*, 29(4):563–574.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Blanco, A., Pino-Mejías, R., Lara, J., and Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from peru. *Expert Systems with applications*, 40(1):356–364.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Boateng, E. Y. and Oduro, F. T. (2018). Predicting microfinance credit default: A study of nsoatreman rural bank, ghana. *Journal of Advances in Mathematics and Computer Science*, pages 1–9.

Bravo, C., Maldonado, S., and Weber, R. (2013). Granting and managing loans for micro-entrepreneurs: New developments and practical experiences. *European Journal of Operational Research*, 227(2):358–366.

Bumacov, V., Ashta, A., and Singh, P. (2014). The use of credit scoring in microfinance institutions and their outreach. *Strategic Change*, 23(7-8):401–413.

Bumacov, V., Ashta, A., and Singh, P. (2017). Credit scoring: A historic recurrence in microfinance. *Strategic Change*, 26(6):543–554.

Calabrese, R., Andreeva, G., and Ansell, J. (2019). "Birds of a feather" fail together: Exploring the nature of dependency in sme defaults. *Risk Analysis*, 39(1):71–84.

Calabrese, R. and Crook, J. (2020). Spatial contagion in mortgage defaults: A spatial dynamic survival model with time and space varying coefficients. *European Journal of Operational Research*, 287(2):749–761.

Carling, K., Jacobson, T., Lindé, J., and Roszbach, K. (2007). Corporate credit risk modeling and the macroeconomy. *Journal of Banking & Finance*, 31(3):845–868.

Cheng, S. (2015). Potential lending discrimination? insights from small business financing and new venture survival. *Journal of Small Business Management*, 53(4):905–923.

Cornée, S. (2019). The relevance of soft information for predicting small business credit default: Evidence from a social bank. *Journal of Small Business Management*, 57(3):699–719.

Crone, S. F. and Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, 28(1):224–238.

Cubiles-De-La-Vega, M.-D., Blanco-Oliver, A., Pino-Mejías, R., and Lara-Rubio, J. (2013). Improving the management of microfinance institutions by using credit scoring models based on statistical learning techniques. *Expert systems with applications*, 40(17):6910–6917.

Cull, R. and Morduch, J. (2018). Microfinance and economic development. In *Handbook of Finance and Development*, pages 550–571. Edward Elgar Publishing Ltd.

Dellien, H. and Schreiner, M. (2005). Credit scoring, banks, and microfinance: balancing high-tech with high-touch. *Microenterprise Development Review*, 8(2).

Dirick, L., Bellotti, T., Claeskens, G., and Baesens, B. (2019). Macro-economic factors in credit risk calculations: including time-varying covariates in mixture cure models. *Journal of Business & Economic Statistics*, 37(1):40–53.

Dorfleitner, G., Just-Marx, S., and Priberny, C. (2017). What drives the repayment of agricultural micro loans? evidence from nicaragua. *The Quarterly Review of Economics and Finance*, 63:89–100.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Fernandes, G. B. and Artes, R. (2016). Spatial dependence in credit risk and its improvement in credit scoring. *European Journal of Operational Research*, 249(2):517–524.

Fox, J. and Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183.

Ganle, J. K., Afriyie, K., and Segbefia, A. Y. (2015). Microcredit: Empowerment and disempowerment of rural women in ghana. *World Development*, 66:335–345.

Geisser, S. (1980). In discussion of g. e. p. box paper entitled: Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A*, 143:416–417.

Gicić, A. and Subasi, A. (2019). Credit scoring for a microcredit data set using the synthetic minority oversampling technique and ensemble classifiers. *Expert Systems*, 36(2):e12363.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77:103–123.

Hoffman, M. D. and Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Imai, K. S., Gaiha, R., Thapa, G., and Annim, S. K. (2012). Microfinance and poverty—a macro perspective. *World development*, 40(8):1675–1689.

Jiang, C., Wang, Z., Wang, R., and Ding, Y. (2018). Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research*, 266(1-2):511–529.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In Flach, P. A., De Bie, T., and Cristianini, N., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer Berlin Heidelberg.

Leroux, B. G., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In Halloran, M. E. and Berry, D., editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 179–191, New York, NY. Springer New York.

Li, J. and Zhenyu, S. (2018). Measuring and managing credit risk for chinese microfinance institutions. *International Journal of Economics and Finance*, 10(7):56–63.

McIntosh, C. (2008). Estimating treatment effects from spatial policy experiments: An application to ugandan microfinance. *Review of Economics and Statistics*, 90(1):15–28.

Pollinger, J. J., Outhwaite, J., and Cordero-guzmán, H. (2007). The question of sustainability for microfinance institutions. *Journal of Small Business Management*, 45(1):23–41.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 71(2):319–392.

Serrano-Cinca, C., Gutiérrez-Nieto, B., and Reyes, N. M. (2016). A social and environmental approach to microfinance credit scoring. *Journal of Cleaner Production*, 112:3504–3513.

Shi, B., Chen, N., and Wang, J. (2016). A credit rating model of microfinance based on fuzzy cluster analysis and fuzzy pattern recognition: empirical evidence from chinese 2,157 small private businesses. *Journal of Intelligent & Fuzzy Systems*, 31(6):3095–3102.

Sohn, S. Y. and Kim, H. S. (2007). Random effects logistic regression model for default prediction of technology credit guarantee fund. *European Journal of Operational Research*, 183(1):427–478.

Spiegelhalter, D. J.and Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B*, 64(4):583–639.

Sun, S. L. and Liang, H. (2021). Globalization and affordability of microfinance. *Journal of Business Venturing*, 36(1):106065.

Thomas, L., Crook, J., and Edelman, D. (2017). *Credit scoring and its applications*. SIAM.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16(2):149–172.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Van Gool, J., Verbeke, W., Sercu, P., and Baesens, B. (2012). Credit scoring for microfinance: is it worth it? *International Journal of Finance & Economics*, 17(2):103–123.

Wang, G., Hao, J., Ma, J., and Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert systems with applications*, 38(1):223–230.

Watanabe (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594.

Zou, K. H., O'Malley, A. J., and Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, 115(5):654–657.

Variable name	Variable description	Effect
Amount	The amount of the loan obtained by the borrower	+
Loan term	The length of time (in months) that the borrower has to repay	NI
Age	The age of the borrower	−
Gender	The gender of the borrower. 0=Male;1=Female	−
n_persons	Household size (number of members)	NI
n_students	Number of students in the household	NI
n_working	Number of labour force in a household	+
n_working_rt	The proportion of labour force in a household	+
Length_resid	Years in the residence	−
Education_n	Dummies for the educational background. 0=College and above; 1=High School; 2= junior middle school; 3=primary school; 4 = no education; Other= No info	+
Poverty	Whether the borrower is a poverty-stricken person.	NI
Property	Whether the borrower is property owner (0) or not (1).	−
House	Whether the borrower owns a house property (0) or not (1).	−
Repayment_n	Dummies for repayment types. 0=Fixed monthly payment at 5% of the principal for 12 month loans; 1=Fixed yearly payment with two months grace period; 2=Fixed yearly payment without grace period	NI
Marital_n	Dummies for marital status. 1= married; 2=divorced; 3=single; 4=widowed and others	NI
Ethnic_n	Dummies for ethnic groups. 0= Han Chinese; 1= Monguor 2= Zhuang people; 3= Tujia people; 4 Hui People; 5= Manchus and other ethnic minorities; Other	NI
Purpose_n	Dummies for loan purpose. 0= Transportation industry; 1= Farming; 2=House reconstruction; 3=Wholesale and retail trades; 4=Social Services; 7=Consumption; 8= Raising livestock and breeding fish; 9=Crop farming; 10=Other	NI
Fi_inst_dep_bal	Per capita deposit balance of financial institutions at city level	NI
Fi_inst_loan_bal	Per capita loan balance of financial institutions at city level	NI
Unempl_rt	Unemployment rate at city level	+

Table 1: Variables used in the credit scoring models.The column 'effect' reports the expected sign of the relationship between a given explanatory variable and the default probability based on the literature (Boateng and Oduro, 2018; McIntosh, 2008; Bellotti and Crook, 2009; Dirick et al., 2019). If we do not have any expectation, we report NI that stands for No Information.

	mean	sd	0.025quant	0.5quant	0.975quant	mode	x_std
(Intercept)	-9.326	1.932	-13.168	-9.309	-5.580	-9.275	
log_amount	0.596	0.175	0.257	0.595	0.944	0.592	0.305
n_students	-0.007	0.104	-0.214	-0.006	0.195	-0.004	-0.005
length_resid	-0.001	0.000	-0.002	-0.001	-0.000	-0.001	-0.194
fin_inst_dep_bal	0.115	0.040	0.041	0.113	0.199	0.110	0.521
fin_inst_loan_bal	-0.208	0.080	-0.378	-0.203	-0.062	-0.194	-0.687
repayment1	-0.557	0.263	-1.094	-0.549	-0.061	-0.535	-0.204
repayment2	-0.421	0.164	-0.744	-0.420	-0.100	-0.419	-0.210
education2	0.133	0.169	-0.194	0.131	0.467	0.128	0.064
education3	0.499	0.287	-0.083	0.506	1.046	0.519	0.136
marital2	1.301	0.192	0.917	1.304	1.671	1.309	0.355
marital3	1.348	0.249	0.845	1.352	1.825	1.360	0.270
ethnicother	1.025	0.294	0.419	1.035	1.573	1.055	0.178
industryno_info	0.484	0.301	-0.142	0.496	1.042	0.520	0.097
gender1	-0.482	0.187	-0.858	-0.479	-0.125	-0.472	-0.215
property1	-0.359	0.260	-0.895	-0.350	0.127	-0.333	-0.116
purpose_loan3	0.242	0.186	-0.131	0.245	0.598	0.251	0.091
purpose_loan7	1.184	0.508	0.110	1.212	2.108	1.267	0.105

Table 2: Parameter estimates for the covariates on individual loans. No random effects. x_std **stands for X-standarisation.**

	mean	sd	0.025quant	0.5quant	0.975quant	mode	x_std
(Intercept)	-12.582	2.151	-16.906	-12.547	-8.459	-12.476	
log_amount	0.690	0.218	0.271	0.687	1.128	0.680	0.236
unempl_rt	0.555	0.134	0.292	0.555	0.820	0.554	0.280
education1	0.653	0.187	0.274	0.658	1.008	0.667	0.170
educationother	0.929	0.473	-0.082	0.958	1.778	1.017	0.077
ethnic1	0.298	0.162	-0.025	0.300	0.610	0.303	0.103
ethnic5	0.380	0.208	-0.043	0.385	0.775	0.394	0.102
ethnicother	0.884	0.300	0.260	0.897	1.439	0.921	0.132
purpose_loan10	0.401	0.219	-0.048	0.408	0.813	0.422	0.094

Table 3: Parameter estimates for the covariates on group loans. No random effects. x_std **stands for X-standarisation.**

	mean	sd	0.025quant	0.5quant	0.975quant	mode
(Intercept)	-10.119	2.131	-14.369	-10.096	-5.999	-10.050
log_amount	0.626	0.193	0.252	0.624	1.010	0.620
n_students	0.028	0.107	-0.184	0.028	0.235	0.030
length_resid	-0.001	0.001	-0.002	-0.001	0.000	-0.001
fin_inst_dep_bal	0.117	0.050	0.028	0.114	0.222	0.108
fin_inst_loan_bal	-0.196	0.098	-0.408	-0.189	-0.024	-0.175
repayment1	-0.410	0.288	-0.993	-0.404	0.139	-0.393
repayment2	-0.430	0.181	-0.787	-0.430	-0.078	-0.428
education2	0.169	0.176	-0.172	0.167	0.517	0.164
education3	0.464	0.305	-0.153	0.470	1.046	0.482
marital2	0.958	0.214	0.529	0.962	1.370	0.968
marital3	1.260	0.264	0.729	1.264	1.767	1.272
ethnicother	0.895	0.332	0.220	0.904	1.522	0.921
industryno_info	0.575	0.315	-0.076	0.586	1.162	0.608
gender1	-0.436	0.193	-0.823	-0.433	-0.067	-0.426
property1	-0.433	0.281	-1.008	-0.425	0.095	-0.408
purpose_loan3	0.277	0.190	-0.104	0.280	0.643	0.285
purpose_loan7	1.155	0.542	0.017	1.180	2.149	1.232
τ	1.248	0.392	0.666	1.185	2.195	1.071

Table 4: Parameter estimates for the covariates on individual loans. Independent random effects.

	mean	sd	0.025quant	0.5quant	0.975quant	mode
(Intercept)	-13.760	2.468	-18.736	-13.714	-9.041	-13.624
log_amount	0.696	0.240	0.237	0.692	1.180	0.683
unempl_rt	0.729	0.237	0.274	0.726	1.204	0.719
education1	0.551	0.195	0.156	0.555	0.922	0.563
educationother	1.043	0.516	-0.048	1.070	1.983	1.125
ethnic1	0.045	0.245	-0.438	0.045	0.524	0.046
ethnic5	0.678	0.270	0.142	0.680	1.202	0.684
ethnicother	0.604	0.361	-0.131	0.614	1.288	0.632
purpose_loan10	0.253	0.234	-0.224	0.259	0.695	0.272
τ	0.733	0.159	0.470	0.716	1.093	0.680

Table 5: Parameter estimates for the variables on group loans. Independent random effects.

	mean	sd	0.025quant	0.5quant	0.975quant	mode
(Intercept)	-9.936	2.121	-14.170	-9.912	-5.838	-9.864
log_amount	0.605	0.192	0.234	0.603	0.988	0.599
n_students	0.024	0.107	-0.188	0.025	0.231	0.026
length_resid	-0.001	0.001	-0.002	-0.001	0.000	-0.001
fin_inst_dep_bal	0.128	0.056	0.029	0.124	0.248	0.117
fin_inst_loan_bal	-0.213	0.109	-0.451	-0.205	-0.023	-0.187
repayment1	-0.327	0.295	-0.922	-0.322	0.238	-0.311
repayment2	-0.444	0.183	-0.806	-0.443	-0.086	-0.442
education2	0.164	0.177	-0.178	0.163	0.514	0.160
education3	0.466	0.305	-0.151	0.472	1.049	0.484
marital2	0.941	0.214	0.511	0.944	1.353	0.950
marital3	1.230	0.265	0.698	1.234	1.739	1.242
ethnicother	0.840	0.336	0.157	0.848	1.477	0.864
industryno_info	0.563	0.316	-0.089	0.574	1.151	0.596
gender1	-0.460	0.193	-0.849	-0.457	-0.090	-0.450
property1	-0.438	0.282	-1.015	-0.430	0.092	-0.413
purpose_loan3	0.279	0.191	-0.103	0.282	0.646	0.288
purpose_loan7	1.079	0.544	-0.061	1.104	2.077	1.155
τ_1	0.528	0.263	0.190	0.472	1.192	0.378
d	1.432	0.869	0.400	1.226	3.676	0.894

Table 6: Parameter estimates for the covariates on individual loans. Spatial random effects with CAR model.

	mean	sd	0.025quant	0.5quant	0.975quant	mode
(Intercept)	-9.984	2.126	-14.228	-9.960	-5.877	-9.913
log_amount	0.610	0.193	0.238	0.608	0.994	0.604
n_students	0.025	0.107	-0.187	0.026	0.233	0.027
length_resid	-0.001	0.001	-0.002	-0.001	0.000	-0.001
fin_inst_dep_bal	0.125	0.054	0.028	0.122	0.242	0.114
fin_inst_loan_bal	-0.209	0.107	-0.441	-0.200	-0.023	-0.183
repayment1	-0.343	0.294	-0.937	-0.338	0.220	-0.328
repayment2	-0.441	0.183	-0.802	-0.440	-0.084	-0.439
education2	0.166	0.176	-0.176	0.164	0.516	0.162
education3	0.465	0.306	-0.152	0.471	1.048	0.483
marital2	0.942	0.214	0.513	0.946	1.355	0.952
marital3	1.236	0.265	0.704	1.240	1.745	1.248
ethnicother	0.852	0.336	0.170	0.860	1.488	0.877
industryno_info	0.565	0.316	-0.087	0.576	1.153	0.598
gender1	-0.455	0.193	-0.844	-0.452	-0.084	-0.445
property1	-0.437	0.282	-1.014	-0.429	0.092	-0.412
purpose_loan3	0.280	0.191	-0.102	0.282	0.646	0.288
purpose_loan7	1.095	0.544	-0.046	1.120	2.093	1.171
τ_2	1.077	0.343	0.563	1.024	1.892	0.927
λ	0.351	0.149	0.100	0.340	0.662	0.309

Table 7: Parameter estimates for the covariates on group loans. Spatial random effects with Leroux model.

	mean	sd	0.025quant	0.5quant	0.975quant	mode
(Intercept)	-13.185	2.501	-18.226	-13.139	-8.401	-13.049
log_amount	0.682	0.242	0.221	0.677	1.170	0.668
unempl_rt	0.622	0.252	0.133	0.619	1.124	0.614
education1	0.528	0.195	0.132	0.532	0.900	0.540
educationother	1.018	0.513	-0.068	1.045	1.952	1.100
ethnic1	-0.044	0.250	-0.535	-0.043	0.445	-0.043
ethnic5	0.668	0.274	0.124	0.669	1.200	0.673
ethnicother	0.484	0.370	-0.267	0.493	1.184	0.510
purpose_loan10	0.239	0.234	-0.239	0.245	0.681	0.258
τ_1	0.223	0.100	0.086	0.204	0.471	0.170
d	2.154	1.148	0.739	1.892	5.104	1.477

Table 8: Parameter estimates for the covariates on individual loans. Spatial random effects with CAR model.

	mean	sd	0.025quant	0.5quant	0.975quant	mode
(Intercept)	-13.361	2.496	-18.392	-13.316	-8.585	-13.226
log_amount	0.686	0.241	0.225	0.681	1.173	0.672
unempl_rt	0.655	0.249	0.172	0.652	1.151	0.648
education1	0.535	0.195	0.139	0.539	0.907	0.547
educationother	1.029	0.515	-0.059	1.056	1.965	1.111
ethnic1	-0.021	0.249	-0.511	-0.021	0.466	-0.020
ethnic5	0.673	0.273	0.131	0.675	1.204	0.678
ethnicother	0.517	0.368	-0.231	0.526	1.215	0.545
purpose_loan10	0.243	0.234	-0.235	0.249	0.685	0.261
τ_2	0.625	0.147	0.383	0.609	0.958	0.579
λ	0.210	0.110	0.054	0.190	0.476	0.143

Table 9: Parameter estimates for the covariates on group loans. Spatial random effects with Leroux model.

	DIC	WAIC	CPO
No RE	1657.887	1659.391	829.716
Independent RE	1554.963	1554.360	777.503
Spatial CAR RE	1549.160	1550.031	775.356
Spatial Leroux RE	1549.635	1550.382	775.529

Table 10: Model comparison for individual loans.

	DIC	WAIC	CPO
No RE	2538.922	2538.639	1269.325
Independent RE	2316.237	2309.684	1155.459
Spatial CAR RE	2320.127	2314.974	1158.012
Spatial Leroux RE	2317.208	2311.828	1156.476

Table 11: Model comparison for group loans.

	H-index	Gini	AUC	KS	Log Score
No RE	0.2196	0.4423	0.7212	0.4100	0.1312
Independent RE	0.2565	0.4834	0.7417	0.4169	0.1261
Spatial CAR RE	0.2646	0.4820	0.7410	0.4013	0.1258
Spatial Leroux RE	0.2615	0.4829	0.7414	0.4054	0.1259

Table 12: Model performance for the out-of-sample (25%) on individual loans.

	H-index	Gini	AUC	KS	Log Score
No RE	0.0966	0.3077	0.6538	0.3124	0.1102
Independent RE	0.2050	0.4557	0.7279	0.3646	0.1041
Spatial CAR RE	0.2173	0.4635	0.7318	0.3562	0.1037
Spatial Leroux RE	0.2129	0.4607	0.7304	0.3594	0.1038

Table 13: Model performance for the out-of-sample (25%) on group loans.

Loan size (log)	Number of loans	Default rate (%)
[7.6,8.85]	32	0.000
(8.85,10.1]	1124	1.068
(10.1,11.4]	6569	3.547
(11.4,12.6]	538	4.089

Table 14: Default rate by loan size for four groups with equal range. Individual data set.

Loan size (log)	Number of loans	Default rate (%)
[6.91,7.89]	26	0.000
(7.89,8.86]	508	0.197
(8.86,9.84]	8831	2.299
(9.84,10.8]	5983	2.708

Table 15: Default rate by loan size for four groups with equal range. Group data set.

link	n_ind	n_grp
0	29	29
1	30	24
2	24	28
3	23	21
4	17	23
5	14	13
6	7	9
7	4	6
8	2	4
9	2	5
10	3	1
11	1	6
12		3
13		2
14		1
15		3
16		1

Table 16: The distribution of the number of links for individual and group loans.

Variable	Statistic	Train	Test
n_students	min	0.00	0.00
	q25	0.00	0.00
	mean	0.73	0.76
	median	1.00	1.00
	q75	1.00	1.00
	max	5.00	5.00
log_amount	min	7.60	8.52
	q25	10.31	10.31
	mean	10.60	10.61
	median	10.82	10.82
	q75	10.82	10.82
	max	12.61	12.61
length_resid	min	12.00	12.00
	q25	336.00	336.00
	mean	423.22	416.48
	median	432.00	424.00
	q75	540.00	540.00
	max	780.00	768.00
fin_inst_loan_bal	min	1.54	1.54
	q25	3.25	3.21
	mean	4.38	4.36
	median	3.51	3.51
	q75	4.69	4.69
	max	30.62	30.62
fin_inst_dep_bal	min	2.80	2.80
	q25	4.74	4.74
	mean	6.67	6.66
	median	5.90	5.60
	q75	7.44	7.44
	max	64.97	64.97
education2	percentage	64.21	62.44
education3	percentage	8.41	7.12
ethnicother	percentage	2.97	3.53
gender1	percentage	26.95	28.12
industryno_info	percentage	4.32	3.97
marital2	percentage	8.08	8.08
marital3	percentage	4.20	4.16
property1	percentage	11.91	11.52
purpose_loan3	percentage	16.80	17.52
purpose_loan7	percentage	0.76	0.92
repayment1	percentage	16.10	15.49
repayment2	percentage	45.54	48.26
default	percentage	3.23	3.24

Table 17: Descriptive statistics for individual data sets, separated by type of sample.

Variable	Statistic	Train	Test
log_amount	min	6.91	6.91
	q25	9.62	9.62
	mean	9.63	9.62
	median	9.62	9.62
	q75	9.90	9.90
	max	10.82	10.82
unempl_rt	min	1.86	1.86
	q25	3.12	3.12
	mean	3.48	3.50
	median	3.53	3.53
	q75	3.90	3.92
	max	4.75	4.75
education1	percentage	7.31	7.35
educationother	percentage	0.78	0.44
ethnic1	percentage	14.20	12.74
ethnic5	percentage	7.60	8.70
ethnicother	percentage	2.30	2.19
purpose_loan10	percentage	6.04	5.32
default	percentage	2.38	2.40

Table 18: Descriptive statistics for group data sets, separated by type of sample.