Spatial dependence in credit risk and its improvement in credit scoring

Guilherme Barreto Fernandes\textsuperscript{a,b,*}, Rinaldo Artes\textsuperscript{a}

\textsuperscript{a} Insper Institute of Education and Research, Rua Quatá, 300, Vila Olímpia, São Paulo, Brazil
\textsuperscript{b} Serasa Experian, Alameda dos Quinimuras, 187, Planalto Paulista, CEP 04068-900 São Paulo, Brazil

\begin{abstract}

Credit scoring models are important tools in the credit granting process. These models measure the credit risk of a prospective client based on idiosyncratic variables and macroeconomic factors. However, small and medium sized enterprises (SMEs) are subject to the effects of the local economy. From a data set with the localization and default information of 9 million Brazilian SMEs, provided by Serasa Experian (the largest Brazilian credit bureau), we propose a measure of the local risk of default based on the application of ordinary kriging. This variable has been included in logistic credit scoring models as an explanatory variable. These models have shown better performance when compared to models without this variable. A gain around 7 percentage points of KS and Gini was observed.

© 2015 Elsevier B.V. and Association of European Operational Research Societies (EURO) within the International Federation of Operational Research Societies (IFORS). All rights reserved.
\end{abstract}

1. Introduction

The correct evaluation of credit risk is an important issue of the Basel agreements. In this context, the probability of default (PD) has a central role. Statistical and mathematical models have been widely employed in order to estimate the PD for companies or contracts. These models, called credit scoring models, usually determine the risk of default conditionally to exogenous factors. The Basel agreements require conservative estimates of PD for loan portfolios, and retail customers – such as small and medium sized enterprises (SMEs) – must be addressed under the perspective of a massive risk evaluation by means of statistical models. In the present paper logistic models (Hosmer & Lemeshow, 2000) will be used to predict the PD of SMEs.

Information on payment history and financial capacity are naturally understood as relevant risk factors in these models. It also seems to be reasonable to assume that the firm location adds information to credit scoring models, particularly those aimed to predict default risk of SMEs. Oftentimes the main customers of these firms are the population and other companies located in the region where they operate. Thus, when considering an SME located in a region that is facing an economic downturn, affecting the performance of nearby businesses, the risk of default of this firm is expected to increase.

In principle, the need of the inclusion of a spatial factor in credit scoring models could be replaced by characteristics of the local economy. However, information gathering is very difficult when the area of investigation is big – once information on small localities in those regions can be rather scarce or unavailable. Similar problems were verified by Gerkman (2011) in a study of real estate prices.

In this context, the analysis of spatial dependence is justified in a comprehensive study on the credit risk of SMEs; few studies on credit scoring, however, consider this effect. The aim of this paper is to incorporate information on default spatial behavior into credit scoring models for SMEs.

The use of an independent ZIP code related variable is a classical alternative to introduce spatial information into credit scoring models. However, it is a qualitative variable with potentially large number of categories, which produces a non-parsimonious model and brings the risk of a multicollinearity problem. Moreover, regions with few individuals would not have good risk assessment. The large number of ZIP-code categories can produce an overfitting effect and may make the model unstable over time. Finally, economic phenomena do not necessarily respect this territorial division.

In this paper, the spatial dependence is considered by the inclusion of a quantitative variable in the model – which may be considered a measure of spatial risk of default – obtained by ordinary kriging (Matheron, 1963). This risk factor is used as an explanatory variable in logistic credit scoring models. Two different alternatives for the inclusion of this factor in the logistic model have been considered. The first, and simplest one, is to consider it as a fixed variable (without measurement error). The other is to admit that the observed value, \( \hat{\tau} \), is, in fact, a proxy of an unobservable variable that expresses the spatial risk factor (\( \tau \)) such that \( \hat{\tau} = \tau + \epsilon \), where \( \epsilon \) is a random error of measurement (logistic model with errors in variables) (Clark, 1982).
To estimate the spatial risk factor, we used a database, provided by Serasa Experian, with information on the default status of all Brazilian SMEs legally established in 2010. A firm is considered an SME if it has annual gross sales of up to 50 million reais (roughly 28 million dollars at that time). The credit scoring model, on the other hand, was applied to a database of a mid-sized bank with operations in Brazil, of approximately 8800 companies that had loans granted between April and June 2010.

The Brazilian banking system includes some of the largest banks in Latin America while it is sound, profitable and well-capitalized (Belaisch (2003)). As stated by Zambaldi, Arana, Lopes, and Politi (2011), the retail market in Brazil is a segment with small loans, high interest rate and decisions based on negative information available in credit bureaus. In Brazil, positive information on payment behavior is limited to the banks the firm has its debts. Usually, Brazilian SMEs keep a relationship with only one financial institution and this creates an information asymmetry to the rest of the banks, even when combined with credit bureau data.

2. Spatial dependence in risk models

The probability of default may be conditioned to various risk factors; the use of the location of the applicant in relation to other borrowers as a potential factor was highlighted by Stine (2011). In this research, the author analyzes the default rates of mortgages in the United States counties. He found evidence of the existence of spatial correlation in the data.

Agarwal, Ambrose, Chomsisengphet, and Sanders (2012) identified the existence of spatial correlation of defaults in mortgage contracts in the United States. However, when other individual characteristics are considered in the models, they concluded that the concentration of sub-prime mortgage in an area does not increase the future credit risk in the neighborhood.

Barro and Basso (2010) suggest a model of contagion that associates the economic relationship of sectors of the economy and the proximity of each pair of firms in a network of firms. Due to computational issues distances between regions were considered instead of firms. The simulation results presented by the authors shed light on the potential results of an existing spatial correlation of default risk of companies.

Unlike Barro and Basso (2010), we use the spatial correlation between companies, rather than between regions. This is also an important difference between our study and Stine’s (2011). While Stine (2011) works on county level, we carry the analysis on considering individual observations of the default status of the companies and their locations in terms of latitude and longitude. Kriging (Matheron, 1963) is a spatial statistics method that can be used at this level of detail.

There are several studies in the literature that incorporate information on spatial distribution of data in studies of the housing market (e.g. Dubin (1992), Bourassa, Cantoni, and Hoelsl (2010), Montero-Lorenzo and Larraz-Iribas (2012), Wong, Yiu, and Chau (2012), Chica-Olmo, Cano-Guervos, and Chica-Olmo (2013)) and studies related to economics and finance (e.g. Agarwal and Hauswald (2010), Bhat, Paleti, and Singh (2012), Gerkman (2011), Giesecke and Weber (2006), Keiler and Eder (2013), Triki and Maktouf (2012)).

Dymski (2006) discusses about the theoretical and empirical research on discrimination and the effect on credit market. The proposed approach measures the effect of the local amount of defaulted companies on SMEs default risk. We do not consider any information on the racial or sex distribution in the neighborhood nor the race or gender of the companies’ owner. In addition, the social characteristics of the neighborhood is not relevant to the research. Another issue is that the methodology presented in this paper does not require a previous identification of areas, what could be interpreted as a way of discriminating populations. Further discussion related to racial discrimination may be found in Scalera and Zazzaro (2001).

2.1. Credit scoring models

According to Thomas, Edelman, and Crook (2002), “Credit Scoring is a set of decision models and their underlying techniques that aid credit lenders in the granting of credit”.

Although credit granting has been around for 4000 years, the concept of credit scoring as we know was developed about 70 years ago. By definition, the purpose of credit scoring models is to identify the profile of good and bad payers, whatever the concept of “good” and “bad” might be. The use of mathematical and statistical techniques for this purpose had its beginnings in the 1940s with Durand (1941), who applied, for the first time, a discriminant analysis to identify good and bad clients. Nevertheless, this model was a research project only, never used as part of a credit worthiness assessment. Only in the 1950s Bill Fair and Earl Isaac founded the first consultancy specialized in credit granting and thereby implemented the first credit scoring in a financial institution. In Brazil, Serasa Experian is one of the main providers of credit scoring solutions since the 1990s (Experian (2014)).

Although the main banks in Brazil use credit scoring systems, the complexity varies from institution to institution. According to Kumar, Nair, Parsons, and Urdapilleta (2006), Caixa (4th largest bank in Brazil) uses a simplified demographic credit scoring, Banco do Brasil (2nd largest bank in Brazil) uses an internally developed credit scoring system based on demographic and behavioral information as well as Serasa credit report (a Brazilian credit bureau).

Thomas et al. (2002) argue that the philosophical motivation behind the study of credit scoring is pragmatism and empiricism. The models are not intended to explain the risk of default, but predict it.

The methodologies used in developing credit-scoring models are based on various mathematical and statistical techniques. Hand and Henley (1997) present a comprehensive review of methods: discriminant analysis (Durand, 1941), ordinary linear regression (Orgler, 1970; 1971), linear programming (Hand (1981), Kolesar and Showers (1985)), regression tree (Makowski (1985)), expert systems (Zocco (1985), Davis (1987)), neural networks (Rosenberg & Gleit, 1994), non-parametric smoothing methods (Hand, 1986) and logistic regression (Wijington, 1980). Thomas (2011) gives an overview of techniques used in credit scoring and concludes that the logistic regression method is by far the most used approach in this sort of modeling.

In the present paper, after the estimation of the regional risk factor, its relevance is tested by its inclusion in a logistic model of credit scoring as an explanatory variable.

3. Methodology

In this section, the kriging method and the logistic regression models used in the analysis are presented.

3.1. Kriging

Kriging is an interpolation method that takes into account the distance between the sampling units and the spatial correlation present in the behavior of the variable of interest. According to Isaaks and Srivastava (1989), ordinary kriging is a method of prediction, via smoothing by means of weighted averages for a new observation. Consider $Z_i$ the value of the variable of interest $Z$ for individual $i$, $i = 1, \ldots, n$. Admit that one wants to predict the value of this variable for a new individual; in this case, the expected value is given by $\hat{Z}_0 = \sum_{i=1}^{n} \lambda_i Z_i$, where $\lambda_i$ is the weight associated with the $i$th individual, subject to $\sum_{i=1}^{n} \lambda_i = 1$. The weights $\lambda_i$ depend primarily on the distance between all observations and the location of the new individual, and they are defined from the spatial dependence structure observed for this variable.

Let $Z_i$ and $Z_j$ be the values of a random variable $Z$ for two subjects placed in a distance $d_{ij} = h$ from each other. Assume that $C(h) =$
The kriging method employed in this work uses the semivariance function - defined as \( \gamma(h) = \frac{1}{2} E[(Z_i - Z_j)^2 | d_{ij} = h] \) as a measure of spatial dependence. It is a measure related to the spatial covariance by \( \gamma(h) = C(0) - C(h) \). The larger the dependence measured by \( C(h) \) is, the smaller the value of the semivariance will be.

\( \gamma(h) \) is estimated by \( \hat{\gamma}(h) = \frac{1}{N_h} \sum_{i,j} (z_i - z_j)^2 \), where \( N_h \) is the number of pairs of observations distant \( h \) units from each other \((i,j)\), \( d_{ij} = h \), and which is an unbiased estimator (Matheron (1963)). The semivariance function graph is named semivariogram, and is an important analysis tool of the spatial dependence present in data (Isaaks & Srivastava, 1989). The components of the semivariogram, represented in Fig. 1, are: (1) the nugget \( (C_0) \), which is the value of the semivariance for \( h = 0 \); (2) the sill \( (C + C_0) \), which represents the limit of the semivariance for \( h \to \infty \); and (3) the range \( (a) \), defined by the value of \( h \) in which the value of the semivariance stabilizes.

To apply kriging, we seek a mathematical model for \( \hat{\gamma}(h) \). There are several theoretical models proposed in the literature, the most common are: Spherical of Matheron (1), the Exponential of Formory (2) and the Gaussian (3). These models are presented below:

1. Spherical Model of Matheron:
\[
\gamma(h) = \begin{cases} 
C_0 + C_1 \left[ 1.5 \left( \frac{h}{a} \right) - 0.5 \left( \frac{h}{a} \right)^3 \right] & \text{for } 0 \leq h \leq a \\
C_0 & \text{for } h > a 
\end{cases}
\]

2. Exponential Model of Formory:
\[
\gamma(h) = C_0 + C_1 \left( 1 - \exp \left( - \frac{h}{a} \right) \right)
\]

3. Gaussian Model:
\[
\gamma(h) = C_0 + C_1 \left( 1 - \exp \left( - \frac{h^2}{\sigma^2} \right) \right)
\]

The semivariance models may be combined to accommodate a more complex spatial correlation structure. According to Hohn (1999), a new model may be obtained from the sum of two simple models:
\[
\gamma_{SUM}(h) = C_0 + \gamma_1(h) + \gamma_2(h),
\]

where \( \gamma_1(h) \) and \( \gamma_2(h) \) are two semivariance models.

The parameter estimates may be obtained by weighted minimum square method (details in Cressie, 1993). The best model to be considered will be the one which produces the minimum weighted sum of squares of errors (WSE), given by WSE = \( \frac{1}{\sum_{i=1}^{n} N_h} \sum_{i,j} (\hat{\gamma}(h) - \hat{\gamma}(h))^2 \), where \( \hat{\gamma}(h) \) is the semivariance predicted by the model, \( H \) is the total number of lags used in the construction of the semivariogram and \( N_h \) is the number of pairs used in the calculation of \( \hat{\gamma}(h) \) (Cressie, 1993).

The estimation of \( Z_0 \) by means of ordinary kriging is given by (5) (Isaaks & Srivastava, 1989):
\[
\hat{Z}_0 = \sum_{i=1}^{n} \lambda_i Z_i, \quad \text{subject to } \sum_{i=1}^{n} \lambda_i = 1
\]

\( \lambda_i \) may be obtained by minimizing the variance of the estimation error, given by \( \min \sigma^2(i) = \min \text{Var}[\hat{Z}_i - Z_i] \), subject to \( \sum_{i=1}^{n} \lambda_i = 1 \).

This solution may be obtained by using Lagrange multipliers. As described in Goovaerts (1997, p. 133), the solution of this optimization problem comes from the solution of (6):
\[
\min \sigma^2(i) = \min \left[ \hat{\sigma}^2 + \sum_{i=1}^{n} \sum_{j=1}^{N} \lambda_i \lambda_j \hat{\gamma}_{ij} - 2 \sum_{i=1}^{N} \lambda_i \hat{\gamma}_{i0} + 2 \mu \left( \sum_{i=1}^{N} \lambda_i - 1 \right) \right].
\]

where \( \hat{\gamma}_{ij} \) is the value of \( \gamma_{ij} \) predicted by the adjusted semivariance model, \( \hat{\gamma}_{i0} \) is the predicted value of the semivariance between point \( i \) and point \( 0 \), \( \mu \) is the Lagrange multiplier and \( \hat{\sigma}^2 \) is the estimated variance of \( Z \).

The solution of (6) is presented in (7) (Isaaks & Srivastava, 1989).
\[
\begin{pmatrix}
\hat{\lambda}_1 \\
\vdots \\
\hat{\lambda}_n \\
\hat{\mu}
\end{pmatrix} =
\begin{pmatrix}
\hat{\gamma}(d_{11}) & \cdots & \hat{\gamma}(d_{1n}) & 1 \\
\vdots & \ddots & \vdots & \vdots \\
\hat{\gamma}(d_{n1}) & \cdots & \hat{\gamma}(d_{nn}) & 1 \\
1 & \cdots & 1 & 0
\end{pmatrix}^{-1}
\begin{pmatrix}
\hat{\gamma}(d_{10}) \\
\vdots \\
\hat{\gamma}(d_{n0})
\end{pmatrix}
\]

In the present case, \( Z_i \) is a binary variable that indicates a default firm with value one and a compliant firm with zero.

The kriging procedure results in a variable \( \hat{Z} \), related to the risk of default of the company due to its location (spatial risk). \( \hat{Z} \) is a real number mostly between zero and one, and, after a capping treatment, all values are restricted to the range of zero to one (logit transformation: logit(X) = ln(X/(1-X))). This is the proxy used to measure the effect of the regional economic factor. The next step will be to test the inclusion of this local risk factor as an explanatory variable in logistic models for credit scoring.

3.2. Logistic regression with errors in the variables

A logistic regression model associates the probability of default with a set of explanatory variables (covariates) by
\[
P(Y_i = 1|X_i = x_i) = p_i = \frac{1}{1 + \exp \left( -x_i \beta \right)}.
\]

where \( Y_i \) is a dummy variable that indicates a default in the contract \( i \), \( X_i \) is the covariate associated with contract \( i \), and \( \beta \) is the parameter vector. The vector \( \beta \) may be estimated by maximum likelihood (Hosmer & Lemeshow, 2000). One of the aims of this paper is to incorporate \( \hat{Z} \) as an explanatory variable in the credit scoring model. We argue that this variable is a proxy of a local spatial risk, so one may consider that \( \hat{Z} = \tau + \epsilon \), in which \( \tau \) represents the real value of the spatial risk and \( \epsilon \) is a zero-mean and homoscedastic random error. In other words, it is measured with error. Model (8) does not incorporate the uncertainty caused by the measurement error. Let \( \hat{\beta}_{\text{NAIVE}} \) be the estimator of \( \beta \) described in (8).

\( \hat{\beta}_{\text{NAIVE}} \) may be asymptotically biased if at least one explanatory variable is measured with error (Clark, 1982). As a consequence of this bias, there is a tendency of underestimation of the probability of default in cases with high risk profiles and of overestimation in cases with low risk profile. Stefanski and Carroll (1985) describe this effect as an attenuation of the estimated probabilities. The method SIMEX (Cook & Stefanski, 1994) was used in the present paper to estimate
the regression model parameters under measurement error assumption. These parameters will be compared with $\hat{\beta}_{\text{NAIVE}}$.

The Kolmogorov–Smirnov (KS) statistics will be used to evaluate the performance of the models (details in Thomas, 2009).

4. Results

In May, 2010, there were 9 million SMEs operating in Brazil according to the Secretariat of the Federal Revenue of Brazil. The kriging method was applied to this database. The default criterion used is similar to the one adopted by the Brazilian Central Bank for retail portfolios (BACEN, 2011): firms that are more than 90 days in arrears on their debts. The data base was provided by the credit bureau Serasa Experian and included the default status of these firms.

Fig. 2 shows the spatial distribution of the firms (random sample of 300 thousand firms) used in the application of the kriging methodology. Note that, the coastal and Southern regions have a greater density of companies and a much lower density all over the Amazon Rainforest area and the Pantanal zone is found. Note that in the northeast of the Brazil’s coast, there is an archipelago called

![Spatial Distribution of Firms](image)

Fig. 2. Firms location.

<table>
<thead>
<tr>
<th>Abbrev.</th>
<th>Region</th>
<th>ZIP code</th>
<th># of SMEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNDSP</td>
<td>São Paulo metropolitan area</td>
<td>01000–000</td>
<td>1,269,923</td>
</tr>
<tr>
<td>INTSP</td>
<td>State of São Paulo, except for GNDSP and LPVP</td>
<td>11000–000</td>
<td>1,248,828</td>
</tr>
<tr>
<td>LPVP</td>
<td>São Paulo coastal cities and Vale do Paraíba</td>
<td>13000–000</td>
<td>284,304</td>
</tr>
<tr>
<td>GNDRJ</td>
<td>Rio de Janeiro metropolitan area</td>
<td>20000–000</td>
<td>357,259</td>
</tr>
<tr>
<td>INTRJ</td>
<td>State of Rio de Janeiro, except for GNDRJ</td>
<td>24000–000</td>
<td>365,702</td>
</tr>
<tr>
<td>ES</td>
<td>State of Espírito Santo</td>
<td>29000–000</td>
<td>168,582</td>
</tr>
<tr>
<td>GNRBH</td>
<td>Belo Horizonte metropolitan area</td>
<td>30000–000</td>
<td>257,975</td>
</tr>
<tr>
<td>INTMG</td>
<td>State of Minas Gerais, countryside</td>
<td>35000–000</td>
<td>693,458</td>
</tr>
<tr>
<td>BA</td>
<td>State of Bahia</td>
<td>40000–000</td>
<td>482,030</td>
</tr>
<tr>
<td>SEAL</td>
<td>States of Sergipe and Alagoas</td>
<td>49000–000</td>
<td>136,853</td>
</tr>
<tr>
<td>PE</td>
<td>State of Pernambuco</td>
<td>50000–000</td>
<td>264,559</td>
</tr>
<tr>
<td>PB</td>
<td>State of Paraíba</td>
<td>58000–000</td>
<td>96,796</td>
</tr>
<tr>
<td>RNCE</td>
<td>States of Rio Grande do Norte and Ceará</td>
<td>59000–000</td>
<td>376,021</td>
</tr>
<tr>
<td>PIMAPA</td>
<td>States of Piauí, Maranhão, Pará and Amapá</td>
<td>64000–000</td>
<td>393,982</td>
</tr>
<tr>
<td>AMMTMS</td>
<td>States of Amazonas, Acre, Mato Grosso do Sul, Mato Grosso, Roraima and Rondônia</td>
<td>78000–000</td>
<td>448,578</td>
</tr>
<tr>
<td>DFGOTO</td>
<td>States of Goiás and Tocantins and the Federal District</td>
<td>70000–000</td>
<td>379,628</td>
</tr>
<tr>
<td>CURPR</td>
<td>Curitiba metropolitan area</td>
<td>89000–000</td>
<td>75,487</td>
</tr>
<tr>
<td>INTPR</td>
<td>State of Paraná, except for CURPR</td>
<td>81000–000</td>
<td>600,356</td>
</tr>
<tr>
<td>SC</td>
<td>State of Santa Catarina</td>
<td>88000–000</td>
<td>424,232</td>
</tr>
<tr>
<td>RS</td>
<td>State of Rio Grande do Sul</td>
<td>90000–000</td>
<td>820,604</td>
</tr>
</tbody>
</table>
The country was stratified in 20 regions due to the regional differences existing in Brazil and to assure better results in the application of kriging methodology. This stratification clustered regions with similar semivariogram behavior and maintained a large number of companies in each stratum. Table 1 brings the distribution of SMEs by region and the related ZIP code area.

The region’s empirical semivariogram is presented in Fig. 3. It should be noticed that the behavior is not homogeneous in the country. The data set used for kriging included the latitude and longitude of each SME. Each unit of distance between two firms is equivalent to 30 kilometers ($h = 1 \Rightarrow 30$ kilometers distance). As shown in Fig. 3, the sill of the empirical semivariogram occurs in a distance inferior to three kilometers ($0.1 h = 3$ kilometers). As shown in Fig. 3, the sill of the empirical semivariogram was reached in a distance inferior to three kilometers. A lack of spatial dependence may be seen in some regions, such as PIMAPA, SC and DFTOGO; in other words, the semivariogram remains almost constant throughout the chart. Fig. 3 shows that for most regions the estimated semivariogram models presented a good fit.

The first step in the analysis of the spatial dependence is the identification of the theoretical semivariogram model that best fits the data. For this, several theoretical models were adjusted and the comparison was carried out by the value of the WSSE statistics. The semivariance models considered in the analysis were the spherical (SPH), the Gaussian (GAU) the exponential (EXP), and models similar to (4) obtained from these three basic models (details in Hohn, 1999).

Table 2 brings the WSSE statistics for each semivariance model adjusted. In order to facilitate the visualization of the results, Table 2 shows the ratio of the WSSE of each model and the maximum WSSE obtained by region (second column of the table). In this table, the best models for each region are highlighted.

The selected models in Table 2 were chosen because: (1) they have the lowest WSSE (to the fourth decimal place) and (2) they are more parsimonious, in that order. The simple Gaussian model was chosen for sixteen of the twenty regions. The semivariance models for other four regions were: PIMAPA with Exponential–Spherical model,
The nugget effect is the measure of the difference between the estimated parameters for different regions, indicating a plurality of spatial dependence. The adjusted semivariogram, which indicates an intermediate spatial correlation at short distances, was used to compare the estimated semivariogram models by region.

<table>
<thead>
<tr>
<th>Region</th>
<th>Maximum WSSE</th>
<th>Ratio: WSSE of the model / maximum WSSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNDRJ</td>
<td>210087.50</td>
<td>1.00 1.00 0.70 0.65 0.54 0.54 0.54 1.00 0.70</td>
</tr>
<tr>
<td>AMMTMS</td>
<td>5951.50</td>
<td>1.00 1.00 0.63 0.63 0.40 0.40 0.40 0.63 0.63</td>
</tr>
<tr>
<td>RS</td>
<td>2044.70</td>
<td>1.00 1.00 0.95 0.95 0.65 0.65 0.65 0.95 0.95</td>
</tr>
<tr>
<td>RNCE</td>
<td>12503.40</td>
<td>1.00 1.00 0.95 0.95 0.58 0.58 0.58 0.95 0.95</td>
</tr>
<tr>
<td>SC</td>
<td>1033.40</td>
<td>1.00 1.00 0.95 0.95 0.58 0.58 0.58 0.95 0.95</td>
</tr>
<tr>
<td>INTPR</td>
<td>19346.60</td>
<td>1.00 1.00 0.95 0.95 0.58 0.58 0.58 0.95 0.95</td>
</tr>
<tr>
<td>GNDSP</td>
<td>742.36</td>
<td>1.00 1.00 0.54 0.54 0.54 0.54 0.54 0.54 0.54</td>
</tr>
<tr>
<td>INTSP</td>
<td>92080.50</td>
<td>1.00 1.00 0.54 0.54 0.54 0.54 0.54 0.54 0.54</td>
</tr>
<tr>
<td>LPVP</td>
<td>29134.50</td>
<td>1.00 1.00 0.54 0.54 0.54 0.54 0.54 0.54 0.54</td>
</tr>
<tr>
<td>CURPR</td>
<td>62089.50</td>
<td>1.00 1.00 0.54 0.54 0.54 0.54 0.54 0.54 0.54</td>
</tr>
<tr>
<td>INTPR</td>
<td>4907.10</td>
<td>1.00 1.00 0.54 0.54 0.54 0.54 0.54 0.54 0.54</td>
</tr>
<tr>
<td>ES</td>
<td>5361.40</td>
<td>1.00 1.00 0.54 0.54 0.54 0.54 0.54 0.54 0.54</td>
</tr>
<tr>
<td>SC</td>
<td>1033.40</td>
<td>1.00 1.00 0.54 0.54 0.54 0.54 0.54 0.54 0.54</td>
</tr>
<tr>
<td>INTSP</td>
<td>92080.50</td>
<td>1.00 1.00 0.54 0.54 0.54 0.54 0.54 0.54 0.54</td>
</tr>
<tr>
<td>LPVP</td>
<td>29134.50</td>
<td>1.00 1.00 0.54 0.54 0.54 0.54 0.54 0.54 0.54</td>
</tr>
<tr>
<td>CURPR</td>
<td>62089.50</td>
<td>1.00 1.00 0.54 0.54 0.54 0.54 0.54 0.54 0.54</td>
</tr>
</tbody>
</table>

The adjusted semivariogram, which indicates an intermediate spatial correlation at short distances, was used to compare the estimated semivariogram models by region.

Table 3 presents the estimated models by region. Note that even though fourteen regions have the same theoretical model, the parameter estimates are different, indicating a plurality of spatial dependency structure among the regions. The nugget effect is the measurement error or random component. The GNDBH with Gaussian–Gaussian model, SC with Spherical model and GNDSP with Gaussian–Spherical model.

Table 3 presents the estimated models by region. Note that even though fourteen regions have the same theoretical model, the parameter estimates are different, indicating a plurality of spatial dependency structure among the regions. The nugget effect is the measurement error or random component. The GNDBH with Gaussian–Gaussian model, SC with Spherical model and GNDSP with Gaussian–Spherical model.

Table 3 presents the estimated models by region. Note that even though fourteen regions have the same theoretical model, the parameter estimates are different, indicating a plurality of spatial dependency structure among the regions. The nugget effect is the measurement error or random component (Cressie, 1993, p. 59). INTPS has the lowest nugget effect among all regions: 45 percent of the sill of the adjusted semivariogram, which indicates an intermediate spatial correlation at short distances. The other areas showed a nugget effect of about 75 to 90 percent of the sill, which means a weaker spatial dependence.

In addition to the spatial credit risk component obtained by ordinary kriging, other factors have influence on the credit risk. These other explanatory variables are observed by the credit bureau within 12 months prior to the granting of credit.

The bank that holds the loan portfolio in this study intended to develop credit scoring models to be applied in two stages: prospection of clients and credit granting. Thus, the predictor variables used in the model are only those available at the credit bureau. (Serasa
The attributes in the model because of industrial secret. A further explanation on generic score from credit bureau is given in (Thomas et al., 2002, p. 16).

Three credit scoring models were adjusted:

- Model 1: logistics model without the incorporation of spatial risk variable obtained by kriging (RISKSPATIAL). This model will serve as a baseline for evaluating the gain that the spatial risk variable adds to the credit scoring performance.
- Model 2: naïve logistics model.
- Model 3: logistics model considering error of measurement for the spatial risk variable (SIMEX method).

Table 4 presents the final adjusted models after the exclusion of non-significant variables.

The use of the model estimated by the SIMEX method (Model 3) did not present relevant differences from Model 2. Such behavior is likely to happen when the measurement error is rather small. In this case, the large dataset used in the estimation of RISKSPATIAL induces a low standard error estimate (one of the parameters in SIMEX method).

In order to measure the influence of each explanatory variable in the model, Altman (2000) proposed the use of a scaled vector that indicates a better performance of Models 2 and 3. However, there are no relevant differences between Model 2 and Model 3, but KS and Gini are almost 7 p.p. lower in Model 1, which indicates a better performance of Models 2 and 3.

### 5. Conclusion

The use of logistic regression models to measure the risk of customer default has been common practice in credit scoring for nearly 20 years. Several explanatory variables have been used: reference file, default history, credit demand, payment history of credit products, among others.

The risk of default of a given company is related to the economic environment in which it finds itself. Small and medium-sized enterprises (SMEs), such as butcher shops, pharmacies and small factories suffer from macroeconomic shocks, but the dependence of the performance of the local economy is a plausible hypothesis. Obtaining economic indexes of small regions, such as cities or neighborhoods, can be an arduous task, especially if the number of such regions is high. In this context, to consider the existence of a latent factor related to the performance of the local economy becomes an interesting alternative to deal with this limitation. In this paper, this factor was obtained from the spatial dependence of the default of small and medium enterprises.

Ordinary kriging (Isaaks & Srivastava, 1989) was used to estimate a spatial risk factor that was used as a latent factor related to the local economy performance. The database used was provided by Serasa Experian and it contains information on the default status of about 9 million small and medium sized Brazilian companies (Fernandes & Araújo, 2012).

Brazil was split into twenty regions and ordinary kriging was applied to each region separately. In most cases, the most suitable
semivariance model for this context is the Gaussian model. The kriging estimate results in the spatial risk variable.

Finally, three credit scoring models have been compared:

- Model 1: logistic credit scoring model without spatial risk variable;
- Model 2: logistic credit scoring model with spatial risk variable as an explanatory variable, and
- Model 3: logistic credit scoring model with spatial risk variable as an explanatory variable measured with error.

Model 1 performed worse than the others, presenting a KS and Gini index around 7 p.p. lower than Models 2 and 3. Furthermore, there were no relevant differences between Models 2 and 3. The hypothesis that the inclusion of a spatial risk factor would improve the performance of a credit scoring model is confirmed once the variable generated to express this characteristic was significant in the models used in this study and that performance indicators (KS and Gini index) showed a better performance when the spatial risk factor was included as an explanatory variable.

In addition, the structure of the spatial dependence is not homogenous in Brazil, but the Gaussian semivariance model seems to fit well in most regions. The main idea behind the spatial risk dependence is the local economic activity that results from the neighborhood relationship with the firms. These connections create a network captured by the distance between businesses. Barro and Basso (2010) explored the idea of connecting firms by their economic sector, using the input-output matrix. Further studies may include the use of other criteria for connecting firms and creating the networks, such as credit inquiries to the bureau or trade receivables record.

In geostatistics, there are several methods used as alternatives to ordinary kriging, such as Gaussian disjunctive kriging, lognormal kriging and indicator kriging (Papritz & Moyeed (1999)). Future research would include a comparison of those methods in a credit risk study, similar to Moyeed and Papritz (2002).

Acknowledgments

This project was partially financed by the Program for the Applied Research from Serasa Experian (Project number: 2012/002). All data used in the paper was supplied by the Analytics team from Serasa Experian.

References


