European Journal of Operational Research 249 (2016) 517–524

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Spatial dependence in credit risk and its improvement in credit scoring

Guilherme Barreto Fernandes[a,b,*], Rinaldo Artes[a]

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

ARTICLE INFO

Article history:
Received 15 March 2014
Accepted 6 July 2015
Available online 29 July 2015

Keywords:
Risk analysis
Spatial dependence
SME credit risk
Ordinary kriging
Credit scoring

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.

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 to 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, $\hat{Z}$, is, in fact, a proxy of an unobservable variable that expresses the spatial risk factor (τ) such that $\hat{Z} = \tau + \varepsilon$, where ε is a random error of measurement (logistic model with errors in variables) (Clark, 1982).

* Corresponding author at: Insper and Serasa Experian Alameda dos Quinimuras, 187 Analytics São Paulo, SP - Brazil. Tel.: +55 11 98642 8017; fax: +55 11 3805 4168.
E-mail address: gbfernandes2002@gmail.com (G.B. Fernandes).

http://dx.doi.org/10.1016/j.ejor.2015.07.013
0377-2217/

To estimate the spatial risk factor, we used a database, provided by Serasa Experian, with information on the default status of all Brazilian SME 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, Aranha, 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 have 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 Hoesli (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 (Makowiski (1985), expert systems (Zocco (1985), Davis (1987)), neural networks (Rosenberg & Gleit, 1994), non-parametric smoothing methods (Hand, 1986) and logistic regression (Wiginton, 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 λ_i is the weight associated with the ith individual, subject to $\sum_{i=1}^{n} \lambda_i = 1$. The weights λ_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 $C(h) =$