The spatial correlation of credit risk and its gain in credit scoring models

Guilherme Barreto Fernandes, MA
Prof Rinaldo Artes, PhD
Credit Cycle

The credit cycle
Credit Risk: probability of default

Related to many risk drivers (SME):
- Indebtness
- Activity sector
- Payment behaviour
- Delinquency history
- Economic information
- Local information

Related somehow in SME risk assessment
Local economic influence in SME

- Economic information
- Local information

Two illustrative explanations

Retailers and other stores

Industrial district
Local economic influence in SME

Retailers and other stores
Local economic influence in SME

Retailers and other stores

Stationery

Butcher shop

Petrol station

Bakery
Local economic influence in SME

Retailers and other stores

- Stationery
- Butcher shop
- Bakery
- Petrol station
Local economic influence in SME

Retailers and other stores

- Stationery
- Butcher shop
- Petrol station
- Bakery
Local economic influence in SME

Retailers and other stores
Local economic influence in SME

Industrial district
Local economic influence in SME

Industrial district

Large Industry

Small Industry 1

Small Industry 2

Small Industry 3
Local economic influence in SME

Industrial district

Small Industry 1
Small Industry 2
Small Industry 3
Local economic influence in SME

Industrial district

Small Industry 1
Small Industry 2
Small Industry 3
Local Economic Factor: Spatial dependence

- Potentially relevant spatial dependence
- How to include this information into credit scoring models?

Local economic indexes

Information gathering difficulties (Gerkman, 2011):
Eg. “Neighbourhood” GDP

Post code grouping

- Easy creation and simple implementation
- Any “Excel-like” software is able to evol such analysis

- Potentially a large number of categories
- Regions with few SME or defaulters may result in poor risk assessment
- May result in unstable model or overfitting issue
Local Economic Factor: Spatial dependence

- Potentially relevant spatial dependence
- How to include this information into credit scoring models?
  1. Local economic indexes
  2. Post code grouping
  3. Out proposal:
     - Spatial dependence effect is a continuous measure

Estimated by Kriging methods
Spatial dependence

✓ Probability of default may be conditioned to many risk factors:
  o Indebtness
  o Reference file
  o Payment behaviour
  o Negative statements info
  o Location

Evidence that location matters (Stine, 2011)
  o Counties level
  o Moran’s I
Spatial dependence
Past work

✓ Argawal et al. (2012) proposed:
  o Neighbourhood information: % of low income people, ethnic mix
  o Not significant in a full model

✓ Barro and Barro (2010):
  o Contagion model
  o Combines location (communa) and industry sector
  o Output: a counterparty PD model

✓ Fernandes (2012):
  o Correlation between firms is conditioned to distance
  o Kriging method estimates SPATIAL RISK
  o Explanatory variable in the credit scoring model

✓ What is Kriging?
Kriging was first used in geology (soil characteristics), epidemiology (risk areas of certain diseases) and agriculture (nutrients concentration).

- Based on the geologist Daniel Krige’s ideas (Krige, 1951)
- Developed by the mathematician Georges Matheron (Matheron, 1963)

✓ Kriging: interpolation method based on distance (Matheron, 1963)
✓ Prediction method via smoothing weighted averages
✓ Ordinary Kriging:

\[
\hat{z}_i = \sum_{j \neq i} \lambda_j^i z_j : \text{average of surroundings of } i
\]

\[
\lambda_j^i = f(d_{ij})
\]

\[
\sum_{j \neq i} \lambda_j^i = 1
\]
Variogram analysis

How does $\lambda_j^i = f(d_{ij})$ changes with $d_{ij}$

Semivariance and semivariogram

$$\gamma(h) = \frac{1}{2} E \left[ (Z_i - Z_j)^2 \mid d(i, j) = h \right]$$

Theoretical variogram model:

Spherical Model of Matheron:

$$\gamma(h) = \begin{cases} c_o + C \left[ 1.5 \left( \frac{h}{a} \right) - 0.5 \left( \frac{h}{a} \right)^3 \right] & , \quad 0 \leq h \leq a \\ c_o + C, & , \quad h > a \end{cases}$$

Exponential Model of Formery:

$$\gamma(h) = c_o + C \left[ 1 - \exp \left( -\frac{h}{a} \right) \right]$$

Gaussian Model:

$$\gamma(h) = c_o + C \left[ 1 - \exp \left( -\frac{h^2}{a^2} \right) \right]$$

Extension (Hohn, 1989):

$$\gamma_{SUM}(h) = c_o + \gamma_1(h) + \gamma_2(h)$$
Variogram analysis

Empirical variogram and forecast

\[ \gamma(h) = \frac{1}{2} E \left[ (Z_i - Z_j)^2 | d(i, j) = h \right] \]

\[ \hat{\gamma}(h) = \frac{1}{2N_h} \sum_{(i,j)|d_{i,j}=h} (Z_i - Z_j)^2 \]

Theorical model estimation
Variogram analysis
Empirical variogram and forecast

How does the semivariogram and semivariance \((\tilde{\gamma}(h))\) connects with \(\hat{Z}_i = \sum_{j \neq i} \lambda_j^i z_j\)?

\[
\min \sigma^2_\epsilon(i) = \min \text{Var}[\hat{Z}_i - Z_i], \text{ subject to } \sum_{i=1}^n \lambda_i = 1.
\]

\[
\begin{pmatrix}
\hat{\lambda}_1 \\
\vdots \\
\hat{\lambda}_n \\
\hat{\mu}
\end{pmatrix} =
\begin{pmatrix}
\tilde{\gamma}(d_{11}) & \cdots & \tilde{\gamma}(d_{1n}) & 1 \\
\vdots & \ddots & \vdots & \vdots \\
\tilde{\gamma}(d_{n1}) & \cdots & \tilde{\gamma}(d_{nn}) & 1 \\
1 & \cdots & 1 & 0
\end{pmatrix}^{-1}
\begin{pmatrix}
\tilde{\gamma}(d_{10}) \\
\vdots \\
\tilde{\gamma}(d_{n0})
\end{pmatrix}
\]

Goovaerts (1997)

New explanatory variable: \(\hat{Z}_i = \text{local average risk for firm } i \) (SPATIALRISK)
Credit Scoring methodology

Credit scoring models

**Naïve logistic regression**

\[ Y_i \sim \text{Bernoulli}(p_i) \]

\[ p_i = \left(\frac{\exp(x'_i \beta)}{1 + \exp(x'_i \beta)}\right) \]

Where:
- \( i = \text{observation} \)
- \( y_i = \text{target variable} \) (1: default; 0: non–default)
- \( p_i = \text{probability of default} \)
- \( x_i = \text{explanatory variables} \)
- \( \beta_{\text{naive}} = \text{parameter vector} \)

- ✔ Estimation via MLE

**Measurement error logistic model**

\[ Y_i \sim \text{Bernoulli}(p_i) \]

\[ p_i = \left(\frac{\exp(x'_i \beta + z_i \theta)}{1 + \exp(x'_i \beta + z_i \theta)}\right) \]

Where:
- \( i = \text{observation} \)
- \( y_i = \text{target variable} \) (1: default; 0: non–default)
- \( p_i = \text{probability of default} \)
- \( x_i = \text{variables without measurement error} \)
- \( z_i = \text{variable with measurement error} \)
- \( \beta = \text{parameters vector} \)
- \( \theta = \text{parameter of the spatial risk variable} \)

- ✔ Estimation via SIMEX (Cook and Stefanski, 1994)
Empirical analysis
Steps for analysis

1. Semivariance and semivariogram estimation
   - Bureau data

2. Estimation of spatial risk
   - Bureau data + specific portfolio

3. Credit scoring estimation
   - Specific portfolio
Empirical analysis

Data used in semivariance estimation

Bureau

9MM SMEs

Information gathered:

✓ Market default: 90 days in arrears
✓ Latitude and longitude
✓ Post code

Spatial dependence pattern may not be equal in every region
### Empirical analysis

#### 20 regions

<table>
<thead>
<tr>
<th>Code</th>
<th>SMEs</th>
<th>Description</th>
<th>SMEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNDSP</td>
<td>1,269,923</td>
<td>Metropolitan areas</td>
<td>1,960,644</td>
</tr>
<tr>
<td>GNDRJ</td>
<td>357,259</td>
<td>States except the metropolitan areas</td>
<td>3,192,648</td>
</tr>
<tr>
<td>GNDBH</td>
<td>257,975</td>
<td>Entire states</td>
<td>2,256,803</td>
</tr>
<tr>
<td>CURPR</td>
<td>75,487</td>
<td>Group of states</td>
<td>1,735,062</td>
</tr>
<tr>
<td>LPVP</td>
<td>284,304</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTSP</td>
<td>1,248,828</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTRJ</td>
<td>365,702</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTMG</td>
<td>693,458</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTPR</td>
<td>600,356</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ES</td>
<td>168,582</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BA</td>
<td>482,030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PE</td>
<td>264,559</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PB</td>
<td>96,796</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SC</td>
<td>424,232</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RS</td>
<td>820,604</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEAL</td>
<td>136,853</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNCE</td>
<td>376,021</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIMAPA</td>
<td>393,982</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMMTMS</td>
<td>448,578</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFGOTO</td>
<td>379,628</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Total SMEs:**
- SMEs: 1,960,644
- States except the metropolitan areas: 3,192,648
- Entire states: 2,256,803
- Group of states: 1,735,062
Empirical analysis
Semivariogram

<table>
<thead>
<tr>
<th>Region</th>
<th>Total number of regions</th>
<th>Strong spatial dependence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metropolitan areas</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>States except the metropolitan areas</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Entire states</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Group of states</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>Semivariogram model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metropolitan areas</td>
<td>Gaussian + Exponential + Spheric</td>
</tr>
<tr>
<td>States except the metropolitan areas</td>
<td>5</td>
</tr>
<tr>
<td>Entire states</td>
<td>5</td>
</tr>
<tr>
<td>Group of states</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>Semivariogram model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metropolitan areas</td>
<td>Gaussian + Exponential + Spheric</td>
</tr>
<tr>
<td>States except the metropolitan areas</td>
<td>5</td>
</tr>
<tr>
<td>Entire states</td>
<td>5</td>
</tr>
<tr>
<td>Group of states</td>
<td>5</td>
</tr>
</tbody>
</table>
Empirical analysis

Spatial risk component

Data

Bureau

Portfolio

Data stratification

Portfolio + spatial risk

Portfolio + 1 variable

Kriging

Kriging

Kriging
Empirical analysis
Credit scoring

- 8,800 firms
- 12 months performance window
- Application scoring model

<table>
<thead>
<tr>
<th>Variables</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naive LR</td>
<td>Naive LR</td>
<td>SIMEX LR</td>
</tr>
<tr>
<td>2 reference file</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>6 credit demand</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>4 default history</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>3 negative statement</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>2 past credit use</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>1 payment capacity</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>1 payment method</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>1 Serasa bureau score</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>Spatial risk component</td>
<td>X</td>
<td>V</td>
<td>V</td>
</tr>
</tbody>
</table>
### Empirical Analysis

#### Credit Scoring

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter estimates</th>
<th>Altman's scaled vector</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
</tr>
<tr>
<td>Intercepto</td>
<td>0.5978*</td>
<td>-0.2563*</td>
</tr>
<tr>
<td>IND_RESTR</td>
<td>-0.1007**</td>
<td>-0.1238*</td>
</tr>
<tr>
<td>RESTR2</td>
<td>-0.3801*</td>
<td>-0.4104*</td>
</tr>
<tr>
<td>RESTR3</td>
<td>-0.3020*</td>
<td>-0.3573*</td>
</tr>
<tr>
<td>SERASA</td>
<td>0.7857*</td>
<td>0.6597*</td>
</tr>
<tr>
<td>DEMCRED6</td>
<td>0.0291*</td>
<td>0.0207*</td>
</tr>
<tr>
<td>DEMCRED1/ DEMCRED2</td>
<td>-0.3434*</td>
<td>-0.3051**</td>
</tr>
<tr>
<td>SPATIALRISK</td>
<td>-</td>
<td>-0.3322*</td>
</tr>
</tbody>
</table>

- 3 negative statement variables
- 3 credit demand variables
- Serasa bureau score
- Bureau score is the most important var.
- Spatial risk is 2nd most important var.
- Model 2 and 3 are much similar
Empirical analysis
Credit scoring

<table>
<thead>
<tr>
<th>Measure</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>KS</td>
<td>29.53%</td>
<td>36.03%</td>
<td>36.02%</td>
</tr>
<tr>
<td>Gini</td>
<td>39.98%</td>
<td>47.13%</td>
<td>47.12%</td>
</tr>
</tbody>
</table>

Hand’s KS/ROC test:

- Model 2 = Model 3
- Model 1 ≠ Model 2
- Model 1 ≠ Model 3
Conclusions

✓ There is evidence from past studies about the spatial dependence

✓ We proposed a proxy for capturing this spatial dependence based on neighbourhood

✓ 20 regions were found to have different spatial dependence patterns

✓ 20 spatial models were estimated for those

✓ The ordinary kriging methodology estimated the proxy for spatial risk

✓ The spatial risk component adds 6.5 p.p. in KS statistic and 7 p.p. in Gini

✓ There was no significant difference found between the naive logistic regression and the logistic model with measurement error


