Quality of life indicators for Italian municipalities: are they risk factors for default probabilities?

Andrea Marletta[1], Caterina Liberati[1]

[1]Università degli Studi di Milano-Bicocca

1. Introduction: Definition, setting, motivation
2. Proposal:
   - Modeling insolvent amount
   - Exploring inter-dependence between spatial dimension and default prediction
3. Case study: Credit risk on Italian SMEs
4. Results: Evidences and Interpretation
5. Summary and Conclusions

Definitions

- A credit risk is the risk of default on a debt that may arise from a borrower failing to make required payments (Basel II)
- The risk is that of the lender and includes lost principal and interest. The loss may be complete or partial
- Three key components:
  1. Expected loss
  2. Variability of the loss on its average value
  3. Diversification effect

Definitions cont'd

Expected loss is composed by:

1. Probability of default (PD): an estimate of the likelihood that a borrower will be unable to meet its debt obligations
2. Exposure at default (EAD): The expected amount of loss that a bank would be exposed to when a debtor defaults on a loan from that bank
3. Loss Given Default (LGD): the share of an asset that is lost if a borrower defaults
4. Maturity: refers to the final payment date of a loan or other financial instrument, at which point the principal (and all remaining interest) is due to be paid.

Modelling Default Prediction of SMEs

- **Credit Risk Models**
  The objective of a Credit Risk model is to develop an accurate rule that can distinguish between good and bad instances (Baesens et al, 2003). Credit risk models can be classified into two groups: 1) models for retail clientele 2) models for corporate sector.

- **Small and Medium Enterprises (SMEs)**
  Small and Medium Enterprises play a central role in the EU economy (Small Business Act of the European Commission 2008)
  SME definition: That is, a business must have an annual turnover of less than 50 million of Euro, a balance sheet total less than 43 million of Euro and the number of employees should not exceed 250 (http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/sme-definition/index.htm)

- **Credit Risk Models for SMEs**
  The overwhelming majority of studies used logistic regression.
  Proportional odds (Fantazzini et al., 2009; Michala et al., 2013),
  Bayesian and classic panel models (Fantazzini et al., 2009)
  Random survival forests (Fantazzini and Figini, 2009)
  Support Vector Machines (Martens et al., 2011)
  BGEVA models (Calabrese and Osmetti, 2013)

Motivation

- Recently, Calabrese et al, (2017) and Mate-Sanchez-Val et al. (2018) have included spatial dimension into a classification model in order to explore if the location could have any inter-dependence with the prediction.

- These works shows how the usage of the spatial variables gives a relevant improvement in the ability to predict default or contagion, although the spatial probit allows to distinguish only between the status: default/no-default. It is not possible to model the insolvency

- Modeling insolvency requests to take into account that the variable is left censored:
  - 0 $\longrightarrow$ amount repayed
  - positive amount $\longrightarrow$ insolvent amount

  Therefore a hard classification model could ignore some information available.

Proposal

- The aim of this work is to provide an exploratory investigation about SMEs default prediction that operate in the South of Italy
- Modeling the insolvent amount using financial and non-financial information that are available from public sources

- The usage of standard risk factors could prevent a lending decision, the standard approach do not yield satisfactory
- Among alternative risk factors the attention has been posed on spatial indicators of the municipalities where monitored SMEs are located

Dataset

- Data were provided by a broker and it refers to Small and Medium Enterprises (SMEs) of Sicily containing information about 5.305 credit lines until 2011
- 189 insolvent loans and 5.116 regular credits

$$insolvency\ rate = \frac{insolvent\ loans}{total\ credits} = \frac{189}{5.305} = 3,6\%$$

The variables considered as potential risk factors for the insolvency state are:

- loan amount
- age of the company
- legal status of the company
- form of financing
- market activity

Aggregations and encoding of variables

Some operations of aggregation and encoding have been applied on original variables to re-arrange data:

- *loan amount*: dichotomized in low (< 20.000 €) and high amount (≥ 20.000 €)
- *age of the company*: categorized in three slots: young (born from 2001 to 2011), medium-age (born from 1991 to 2000) and old companies (born before 1991)
- *legal status of the company*: refers to the sole trader and the companies
- *form of financing*: distinguished between defined expiration of the loan (mortgages) and non-defined loan (bank overdraft facilities)
- *market activity*: divided into three macro-categories: agricultural, industry and services

Preliminary descriptive statistics

Risk factor	Frequency	Percentage (%)
Loan amount		
$0 - 20.000$ €	2.102	$39,6\%$
$20.000+$ €	3.203	$60,4\%$
Age of the company		
Young	2.638	$49,7\%$
Medium-age	1.779	$33,6\%$
Old	888	$16,7\%$
Legal status		
Sole trader	3.030	$57,1\%$
Companies	2.275	$42,9\%$
Form of financing		
Bank overdraft	2.221	$41,9\%$
Mortgage	3.084	$58,1\%$
Market activity		
Agriculture	142	$2,7\%$
Services	4.280	$80,7\%$
Industry	883	$16,6\%$

Censored regression: Tobit Model I

In this work, since dependent variable is represented by the default loan and the insolvency rate is $3,6\%$, it is a left-censored variable with many 0 values.

A censored model is based on the idea of a latent, or unobserved variable that is not censored, and is explained via a probit model.

$$y_i^{\star} = \beta_1 + \beta_2 x_i + e_i$$

The observable variable, y, is zero for all $y^{\star}$ that are less or equal to zero and is equal to $y^{\star}$ when $y^{\star}$ is greater than zero. The model for censored data is called Tobit I (Tobin, 1958).

$$y = \begin{cases} 0 & y_i^{\star} \leq 0 \\ y_i^{\star} & y_i^{\star} > 0 \end{cases}$$

$$P(y=0) = P(y^{\star} \leq 0) = 1 - \Phi[(\beta_1 + \beta_2 x)/\sigma]$$

Censored regression: Tobit Model II

Tobit model I are useful when the sample selection is not random, but whehter an individual is in the sample depends on individual characteristics.

The model to use in such situation is Heckit (Heckman, 1979) or Tobit model II (Amemiya, 1985), which involves two equations:

1. selection equation: $z_i^* = \gamma_1 + \gamma_2 w_i + u_i$
2. regression equation: $y_i = \beta_1 + \beta_2 x_i + e_i$

Estimates of the β_s can be obtained by using least squares on the model

$$y_i = \beta_1 + \beta_2 x_i + \beta_\lambda \lambda_i + \nu_i \qquad\qquad \lambda_i = \frac{\phi(\gamma_1 + \gamma_2 w_i)}{\Phi(\gamma_1 + \gamma_2 w_i)}$$

The λ_i is called the inverse Mills ratio, the Heckit procedure involves two steps, estimating both the selection equation and the equation of interest.

Tobit I model: first attempt

- Results for Tobit I model for Default amount are not satisfactory
- Only intercept and loan amount coefficient are significant
- Necessity to include other variables related to location

	Dependent variable:
	Default amount
Loan amount	$-4{,}806.824^{**}$
	(2,210.121)
Legal Status	1,025.820
	(2,198.037)
Form of financing	2,478.243
	(2,224.171)
Age of the company (1990-2000)	3,679.619
	(3,207.361)
Age of the company (2001-2010)	2,320.303
	(3,381.508)
Market activity Agriculture	-509.712
	(6,860.706)
Market activity Industry	3,345.979
	(2,786.787)
logSigma	10.410^{***}
	(0.063)
Constant	$-62{,}742.110^{***}$
	(5,249.252)
Observations	5,305
Log Likelihood	$-2{,}768.391$
Akaike Inf. Crit.	5,554.783
Bayesian Inf. Crit.	5,613.970
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

How do exploit information about municipalities?

Location of SMEs is available for the analyzed dataset. This information could be used creating 3 new variables:

1. A dummy variable indicating whether the municipality is a province
2. A dummy variable indicating the province of Palermo
3. Territorial indicators obtained via Principal Component Analysis

Variables selected for building latent factors are:

- Average income (in euros, source: Agenzia Entrate)
- Occupation rate (in %, source: Istat)
- Education level (in %, source: Istat)
- Population density (in Pop. per km^2, source: Istat)
- Altitude (in meter, source: Istat)
- Euro4+ cars rate (in %, source: ACI)
- Participation al last elections (in %, source: Ministry of the Interior)

Territorial indicators

- 77% of the variance is explained by 3 components:

Total Variance Explained

Component	Initial Eigenvalues Total	Initial Eigenvalues % of Variance	Initial Eigenvalues Cumulative %	Extraction Sums of Squared Loadings Total	Extraction Sums of Squared Loadings % of Variance	Extraction Sums of Squared Loadings Cumulative %
1	3,365	48,076	48,076	3,365	48,076	48,076
2	1,334	19,052	67,128	1,334	19,052	67,128
3	0,681	9,724	76,853	0,681	9,724	76,853
4	0,592	8,463	85,316			
5	0,559	7,984	93,300			
6	0,354	5,057	98,357			
7	0,115	1,643	100,000			

Extraction Method: Principal Component Analysis.

1. Human capital
2. Well-being
3. Physical features

Rotated Component Matrix[a]

	Component 1	Component 2	Component 3
Density	0,012	0,901	0,076
Altitude	0,016	-0,199	-0,874
Income	0,506	0,664	0,378
Education	0,675	0,476	0,319
Occupation	0,853	-0,020	0,151
Elections	0,723	0,201	-0,349
Euro.4 Cars	0,270	0,599	0,518

Extraction Method: Principal Component Analysis.
a. Rotation converged in 6 iterations.

Model with territorial variables

	Dependent variable:		
	Default amount		Yes/No
	Tobit I	*Tobit I/Sigma*	*Probit*
Dummy Loan Amount	$-4{,}804.479^{**}$	-0.148	-0.234^{***}
	(2,187.722)		(0.067)
Prov.Palermo	$-6{,}215.978^{**}$	-0.191	-0.196^{**}
	(2,940.970)		(0.090)
No Countyseat	$7{,}375.540^{*}$	0.227	0.265^{**}
	(4,219.519)		(0.129)
Ind.Humancapital	$-4{,}317.668^{***}$	-0.133	-0.137^{***}
	(1,599.864)		(0.047)
Ind.Wellbeing	$2{,}805.721^{***}$	0.086	0.094^{***}
	(1,000.468)		(0.030)
Ind.Physical	2,169.423	0.067	0.079
	(1,667.133)		(0.052)
logSigma	10.389^{***}		
	(0.062)		
Constant	$-67{,}556.200^{***}$	-2.079	-2.070^{***}
	(6,464.303)		(0.155)
Observations	5,305	5,305	5,305
Log Likelihood	$-2{,}752.606$		-788.237
Akaike Inf. Crit.	5,521.212		1,590.473
Bayesian Inf. Crit.	5,573.823		
Note:			*p<0.1; **p<0.05; ***p<0.01

Interpretation of the coefficients

Relation with Default amount		
Positive	**Negative**	**Not-significant**
No Countyseat	Dummy Loan Amount	Physical features
Well-being	Prov.Palermo	
	Human capital	

- Small loans are more at risk
- SMEs located in Palermo area are less at risk
- SMEs located in small municipalities are more at risk
- SMEs located in municipalities with high occupation and education level are less at risk
- SMEs located in municipalities with higher income are more at risk (see European Parliament, 2015 for Sicilian economy)

Tobit II - Heckman model

Tobit II

	Dependent variable:	
	log(Default amount)	
	Heckman 2-step	Heckman ML
Loan amount	0.00001^{***}	0.00001^{***}
	(0.00000)	(0.00000)
Prov.Palermo	-0.600^{*}	-0.147
	(0.362)	(0.195)
No Countyseat	1.010^{**}	0.541^{**}
	(0.510)	(0.275)
Ind.Humancapital	-0.372	-0.028
	(0.233)	(0.130)
Ind.Wellbeing	0.293^{**}	0.115
	(0.138)	(0.072)
Ind.Physical	0.176	0.025
	(0.191)	(0.118)
Constant	0.063	6.461^{***}
	(2.933)	(0.679)
Observations	5,305	5,305
Log Likelihood	$-1{,}022.331$	
ρ	1.038	0.625^{***} (0.153)
Inverse Mills Ratio	3.283^{***} (1.153)	
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Further considerations about Tobit II

- For this extension of the censored model, the dependent variable is now the logarithm of default amount for the regression equation, while loan amount is now considered as continue explicative variable.
- The λ_i inverse Mills ratio is positive and significant different from 0 justifying the use of this model
- Differently from Calabrese et al. (2017), a geographic dependence is evidenced while structural variables of the company such as legal form, industry sector, and number of employees seem to not affect the suffered loan.
- As in Mate-Sanchez-Val et al. (2018), a territorial component is present but with inverse results about distance from the city centre
- This could due to a different structure in SMEs between Sicily and metropolitan areas (London and Madrid), considering for example the dimension of the companies in terms of number of employees and revenues

Conclusions and future works

- The proposed approach combines the use of censored regression with territorial indicators for measuring credit risk in SMEs context
- Availability of information about default amount for credit lines suggested the use of Tobit model I and II
- Selected models showed a significant association of the territorial indicators based on ACP
- The risk of default is higher for municipalities with a lower human capital and higher well-being
- Future works could imply the use of a spatial correlation matrix between municipalities and financial indexes of the companies