Spatial regression models to improve P2P credit risk management

Arianna Agosto, Paolo Giudici and Tom Leach
University of Pavia

Abstract

Calabrese et al. (2017) have shown how binary spatial regression models can be exploited to measure contagion effects in credit risk arising from bank failures. To illustrate their methodology, the authors have employed the Bank for International Settlements' data on flows between country banking systems. Here we apply a binary spatial regression model to measure contagion effects arising from corporate failures. To derive interconnectedness measures, we use the World Input-Output Trade (WIOT) statistics between economic sectors. Our application is based on a sample of 1185 Italian companies. We provide evidence of high levels of contagion risk, which increases the individual credit risk of each company.

Keywords: Credit Risk, Systemic Risk, Contagion, Spatial Autoregressive models, binary data.

1 Introduction

In recent years, the emergence of financial technologies (fintechs) is redefining the roles of financial intermediaries and introducing many opportunities for consumers and investors. In particular, peer-to-peer (P2P) online lending platforms allow private individuals to directly make small and unsecured loans to private borrowers.
P2P lending business models vary in scope and structure: a comprehensive review is provided by Claessens et al. (2018). Here we specifically refer to the platforms that lend to small and medium enterprises (SME).
While both classic banks and P2P platforms rely on credit scoring models for the purpose of estimating the credit risk of their loans, the incentive for model accuracy may differ significantly. In a bank, credit risk assessment is conducted by the financial institution itself which, being the actual entity that assumes the risk, is interested to have the most accurate possible model. In a P2P lending platform, credit risk is determined by the platform but the risk is fully borne by the lender. In other words, P2P platforms allow for direct matching between borrowers and lenders.
A factor that penalizes the accuracy of P2P credit scoring models is that they often do not have access to borrowers' data usually employed by banks, such as account transaction data, financial data and credit bureau data.
For these reasons, the accuracy of credit risk estimates provided by P2P lenders may be poor.

However, P2P platforms involve their users and, in particular, the borrowers, in a continuous networking activity. Data from such activity can be leveraged not only for commercial purposes, as it is customarily done, but also to improve credit risk accuracy.
We believe that networking information can offset the lack of financial and credit behavioural data and improve credit risk measurement accuracy of P2P lenders, but also of banks. There are indeed cases in which also traditional financial intermediaries face lack of information about the borrower. Consider, for example, credit granting to new customers, for whom internal behavioural data - known to be the most predictive in rating models - are not available.
When financial networks are backed by statistical models, inferential statements can be obtained. Important contributions in this framework are Billio et al.(2012); Diebold and Yilmaz, (2014), Hautsch et al.(2016), Ahelegbey et al. (2016), Giudici and Spelta (2016), Giudici and Parisi (2018), who propose measures of connectedness based on similarities, Granger-causality tests, variance decompositions and partial correlations between market price variables. We improve these contributions, extending them to the P2P context and linking network models, that are often merely descriptive, with econometric models, thus providing a predictive framework. More specifically, we suggest to use spatial econometrics to study the interconnectedness in the corporate sector. Spatial econometrics incorporates dependence among observations that are in any kind of proximity, not only geographical. In particular, the model we apply is a logit Spatial Autoregressive model based on an exogenously defined network. The main advantages of this approach over the traditional network analysis is that it can be used as both an early warning model, to forecast the failure of a given company, and as a stress testing technique taking systemic effects into account.

The paper is organized as follows. Section 2 explains the econometric methodology. Section 3 presents the results obtained applying the proposed methodology to data collected from a European P2P lending information provider. Section 4 concludes.

2 Methodology

2.1 Spatial logit

The model we use in this paper has a binary spatial autoregressive structure, whereby the dependent variable is binary and a spatial autoregressive structure is assumed in the underlying latent variable. Taking the latent underlying quantity to be represented by a continuous variable y_i^*, we consider the observation mechanism as

$$y_i = \begin{cases} 1, & y_i^* > 0 \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

with $i = 1, 2, \ldots, n$.

We implement the spatial structure with an autoregressive model specification, such that

$$Y^* = \rho W Y^* + X\beta + \varepsilon, \tag{2}$$

where Y^* is a continuous random vector, X represents an $n \times k$ matrix of explanatory variables, ε is the error term and W is the spatial lag weight matrix with ρ the associated coefficient, which in our application to defaults will be interpreted as a contagion parameter.

The model implies heteroskedastic errors e as follows:

$$Y^* = (I - \rho W)^{-1}(X\beta + \varepsilon) = (I - \rho W)^{-1}X\beta + e, \tag{3}$$

where

$$e = (I - \rho W)^{-1}\varepsilon \tag{4}$$

and

$$var(e) = var\left[(I - \rho W)^{-1}\varepsilon\right] = \sigma_\varepsilon^2\left[(I - \rho W)'(I - \rho W)\right]^{-1}. \tag{5}$$

The defined model has been used by Calabrese et al. (2017) to study default interdependence in the European banking sector.
Relative to the estimation, Calabrese and Elkink (2014) have provided a review of the main methodologies for model (3) in the literature.
Among the various approaches, we focus on the Generalised Method of Moments (GMM) proposed by Pinkse and Slade (1998). They derive the Generalised Method of Moments (GMM) moment equations from the likelihood function of a spatial error probit model, for which Klier and McMillen (2008) provide the extension to logit models. The GMM approach does not rely on a potentially inaccurate assumption of normally distributed errors and is therefore more robust than maximum likelihood methods.

In general, a GMM estimator is defined by:

$$\hat{\theta} \equiv \arg\min_{\Theta} m_n(\theta)\Omega_n m_n(\theta)', \tag{6}$$

where $m_n(\theta)$ are the moment conditions and Ω_n is a weighting matrix to be determined.
In our case, we have:

$$\theta = [\rho, \beta]$$

To construct the moments, following Pinske and Slade (1998) we use the generalised residuals

$$u_i = y_i - p_i, \tag{7}$$

where:

$$p_i = Pr[y_i = 1] = \frac{\exp^{(I-\hat{\rho}W)^{-1}X\hat{\beta}}}{1 + \exp^{(I-\hat{\rho}W)^{-1}X\hat{\beta}}}$$

It follows from specification (3) that the elements of the spatially lagged dependent vector WY^* are correlated with those of the error vector, hence the need for instrumental variables. Following Kelejian and Prucha (1998), who suggest to choose the instruments as a subset of the linearly independent columns of:

$$H = \{X, WX, W^2X, W^3X, ...\}$$

,

we define the instrument matrix [1]

$$Z = \{X, WX\}$$

Thus, generating the moment conditions via the identity:

$$E[Z'u] = 0$$

$\hat{\theta}$ can be estimated by the following

$$\hat{\theta} = \arg\min_{\Theta} u'Z\Omega Z'u \tag{8}$$

The estimation algorithm used in our application is explained in detail in Section 2.3.

2.2 The network

The spatial regression model we propose is based on an exogenously defined network, where the nodes correspond to individual companies and the ties express the volume of trade between any pair of companies, i.e. the trade flow from company i to company j, for each i and each j. This information is generally not available, so we must approximate it using data on aggregate input-output trade between sectors.

The World Input Output Trade (WIOT) statistics provide information on the aggregate trade volumes of 52 economic sectors in each country with all sectors in all countries.

For a given country, define A as the sector of company i, B as the sector of company j, and let f_{AB} be the trade flow from sector A to sector B, while f_{BA} is the trade flow from sector B to sector A.

Replacing the individual flows with the aggregate ones, the entries of the approximate trade matrix F are then obtained as:

$$f_{ij} = f_{AB} = \sum_{l \in A} \sum_{m \in B} f_{lm}$$

To use these data for proxying the individual companies' flows, we need to calculate the proportion of each company in terms of size over its sector using a suitable measure, such as turnover or the value of trade receivables (for inflows) and payables (for outflows).
Consider, for example, the case of determining the trade flows from company i, belonging to sector A, to company j, belonging to sector B, knowing the individual trade payables and receivables.
We first calculate the ratio between company i trade payables and the sum of sector A trade payables:

$$\bar{x} = \frac{x_i}{\sum_{l \in A} x_l}$$

[1] As explained in Kelejian and Prucha (2010), H proxies the expected value of WY^* using its projection on X.

Then we calculate the ratio between company j trade receivables and the sum of sector B trade receivables:

$$\bar{y} = \frac{y_i}{\sum_{m \in B} y_m}$$

The product $\bar{x}\bar{y}$ is a proxy of the proportion of flows from company i to company j on the total flows from sector A to sector B.
Repeating this calculation for all companies, we get the matrix:

$$R = \langle \bar{x}, \bar{y} \rangle = \begin{pmatrix} \bar{x}_1\bar{y}_1 & \bar{x}_1\bar{y}_2 & \cdots & \bar{x}_1\bar{y}_n \\ \bar{x}_2\bar{y}_1 & \bar{x}_2\bar{y}_2 & \cdots & \bar{x}_2\bar{y}_n \\ \vdots & \ddots & \cdots & \vdots \\ \bar{x}_n\bar{y}_1 & \bar{x}_n\bar{y}_2 & \cdots & \bar{x}_n\bar{y}_n \end{pmatrix}$$

Finally, by calculating the entrywise product of R and the trade matrix F, we get the following matrix:

$$W = R \circ F = \begin{pmatrix} \bar{x}_1\bar{y}_1 R_{1,1} & \bar{x}_1\bar{y}_2 R_{1,2} & \cdots & \bar{x}_1\bar{y}_n R_{1,n} \\ \bar{x}_2\bar{y}_1 R_{2,1} & \bar{x}_2\bar{y}_2 R_{2,2} & \cdots & \bar{x}_2\bar{y}_n R_{2,n} \\ \vdots & \ddots & \cdots & \vdots \\ \bar{x}_n\bar{y}_1 R_{n,1} & \bar{x}_n\bar{y}_2 R_{n,2} & \cdots & \bar{x}_n\bar{y}_n R_{n,n} \end{pmatrix}$$

Note that the ij element can be interpreted as the proxy of the trade flow from company i to company j. Conversely, the ji element can be interpreted as the proxy of the trade flow from company j to company i. The estimated flows define the magnitude of intercompany connections. To use W as a spatial weighting matrix in our application, we need to set the entries on the diagonal to 0 and normalize the rows so as to sum to 1.

2.3 Estimation procedure

To estimate the SAR model parameters, we use a two-step estimation procedure:

i) minimize equation (8), letting $\Omega = I_{3k}$, to obtain parameter estimates $\hat{\theta}$ and calculate the optimal weighting matrix by computing the covariance of the moments:

$$\hat{S} = \frac{1}{n} u' Z Z u'$$

where the residual vector u is calculated as in (7).

ii) recompute the parameter estimates $\hat{\theta}$ by substituting the identity matrix with the optimal weight matrix:

$$\hat{\Omega} = \hat{S}^{-1}$$

Note that this procedure requires inversion and multiplication of large matrices, so the computation time can be very long when working with large datasets. Possible solutions should be based on suitable simplifications to the connectivity matrix W to make it more sparse, such as fixing a threshold for the relevance of trade flows. However, with our sample size ($n = 1185$) the computational time for the two-step algorithm is more than acceptable.

3 Data and results

In this section we empirically verify whether the predictive performance of P2P credit scoring models can be improved using correlation network models. In particular, we are interested in assessing significance and magnitude of the contagion parameter ρ. The more the contagion parameter is close to 1, the more the networking information can support credit risk evaluation. To achieve this goal, we have collected data from a European Credit Assessment Institution (ECAI), that supplies credit scorings to P2P platforms specialised in business lending.
We use data relative to 1185 borrowing Italian SMEs, in 2015-2016. The proportion of observed defaults in our sample is nearly 11%, which is large, in line with the observed impact of the recent financial crisis in Southern European countries.
The available data include the status of the companies, classified as [1 = Defaulted] and [0 = Active], in 2016 as well as some main financial information, for year 2015.
From the available data, we select three financial ratios reflecting the three most important aspects related to default probability: operational performance, business sustainability and financial sustainability. Specifically, we consider:

- the return on equity ratio (RATIO012)
- the activity ratio, expressed as the ratio between sales and total assets (RATIO018);
- the solvency ratio, expressed as the ratio between the net income and the total debt (RATIO027)

The spatial weight matrix W has been built from the WIOT database, as described in Subsection 2.2 and using turnover as a company size measure. Figure 4 shows the network based on the estimated connections.

Figure 1 about here

Table 1 shows the parameter estimates obtained using a simple logit model, without the spatial component.

Table 1 about here

Then we estimate the SAR model (3) through the algorithm presented in Section 2.3. The obtained results are reported in Table 2.

Table 2 about here

We first note from Table 2 than the contagion parameter is significant and its value is high (0.78). The effect of financial ratios is stable, supporting the SAR specification including both a spatial and an exogenous component. Thus, considering a measure of connectivity between companies significantly explains the credit risk arising from P2P lending, improving the traditional analysis based on individual financial indicators.
Including the spatial component also improves model accuracy, as shown in Figure 4 plotting the ROC curves of the simple logit and the spatial logit model. The AUC (Area Under the ROC Curve) values are 0.798 and 0.806 respectively.

Figure 2 about here

4 Conclusions

This paper provides a method, based on binary spatial regression models, to improve default prediction by estimating the interdependence between companies due to trade ties.

We have applied the methodology to a sample of Italian companies, finding evidence of a high level of spatial autocorrelation, interpretable as a credit contagion parameter.

The proposed model provides both a description of contagion (through the spatial component) and a predictive capability, differently from most existing contagion models, which provide either of the two. The model can be easily implemented, as a modification of a classical logistic regression that includes interconnectedness. We believe that the findings which can be derived from spatial autoregressive models may be useful, especially for P2P lenders who can use it to improve credit risk assessment.

From a methodological viewpoint, further research may involve employing a different generalised linear model, such as the generalized extreme value regression models discussed in Calabrese and Elkink (2016). Moreover, the dependence structure could be extended to the dynamic case (Arakelian and Dellaportas, 2012).

References

[1] Ahelegbey, D.F., Billio, M. and Casarin, R. (2016). Bayesian Graphical Models for Structural Vector Autoregressive Processes, Journal of Applied Econometrics 31(2), 357-386.

[2] Arakelian, V. and Dellaportas, P. (2012). Contagion determination via copula and volatility threshold models, Quantitative finance 12(2), 295-310.

[3] Billio, M., Getmansky, M., Lo, A.W. and Pelizzon, L. (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. Journal of Financial Economics, 104(3), 535-559.

[4] Calabrese, R. and Elkink, J. (2014). Estimators of binary spatial autoregressive models: A Monte Carlo study. Journal of Regional Science 54(4), 664-687.

[5] Calabrese, R. and Elkink, J. (2016). Estimating Binary Spatial Autoregressive Models for Rare Events. Advances in Econometrics 37, 147-168.

[6] Calabrese, R., Elkink, J. and Giudici, P. (2017). Measuring Bank Contagion in Europe Using Binary Spatial Regression Models . Journal of the Operational Research Society 68(12), 1503-1511.

[7] Claessens, S., Frost, J., Turner, G. and Zhu F. (2018). Fintech Credit Markets around the World: Size, Drivers and Policy Issues, BIS Quarterly Review, September 2018.

[8] Diebold, F.X. and Yilmaz, K. (2014) On the network topology of variance decompositions: Measuring the connectedness of financial firms. Journal of Econometrics, 182(1), 119-134.

[9] Giudici, P. and Parisi, L. (2018). CoRisk: Credit Risk Contagion with Correlation Network Models. Risks 2018, 6(3), 95.

[10] Giudici, P. and Spelta, A. (2016). Graphical network models for international financial flows. Journal of Business and Economic Statistics 34(1), 126-138.

[11] Hautsch, N., Schaumburg, J. and Schienle, M. (2015). Financial Network Systemic Risk Contributions. Review of Finance, 19(2), pp. 685-738.

[12] Kelijian, H.H. and Prucha, I.R. (1998). A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model With Autoregressive Disturbances. Journal of Real Estate Finance and Economics, 17(1), 99-121.

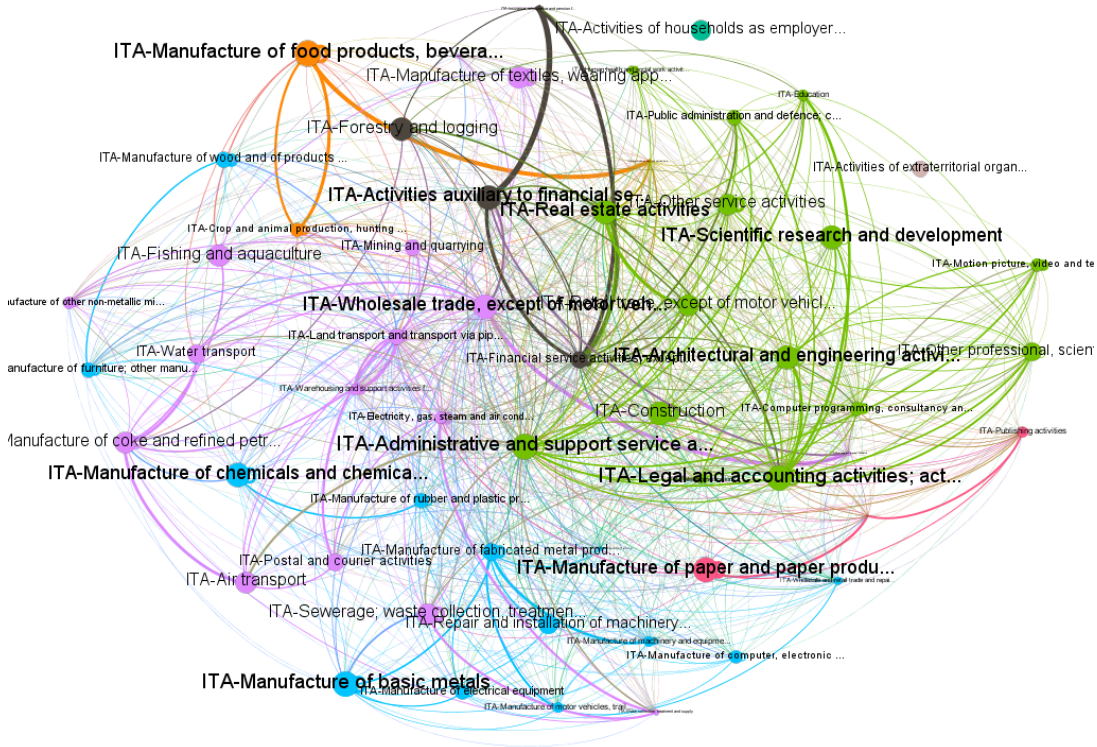
[13] Kelijian, H.H. and Prucha, I.R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. Journal of Econometrics, 157(1), 53-67.

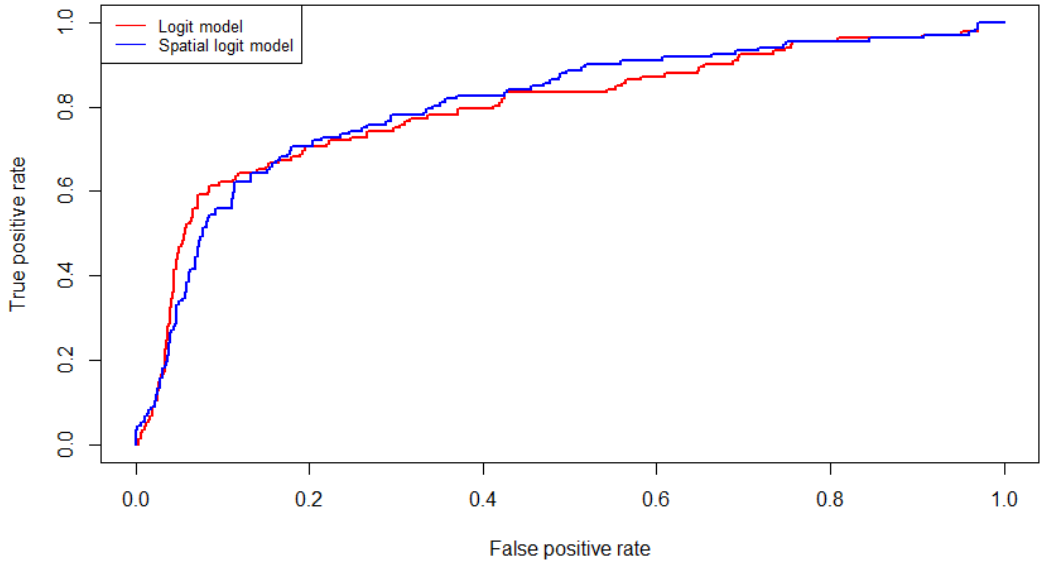
[14] Klier, T. and McMillen, D.(2008). Clustering of auto supplier plants in the United States: generalized method of moments spatial logit for large samples. Journal of Business and Economic Statistics 26(4), 460-471.

[15] LeSage, J. and Pace, R.K (2009). Introduction to Spatial Econometrics. CRC Press.

[16] Pinkse, J. and Slade, M.E. (1998). Contracting in space: an application of spatial statistics to discrete-choice models. Journal of Econometrics 85(1), 125-154.

Figures and Tables





	Estimate	Std. Error	Pr($>|z|$)
Intercept	-2.11	0.16	2.97e-38
β_1 (RATIO012)	-0.69	0.10	6.35e-11
β_2 (RATIO018)	0.02	0.10	0.84
β_3 (RATIO027)	-0.01	0.00	9.10e-04

Table 1: Results of estimation of non-spatial logit model

	Estimate	Std. Error	Pr($>|z|$)
ρ	0.78	0.23	5.44e-04
Intercept	0.44	0.46	0.35
β_1 (RATIO012)	-0.53	0.15	2.24e-04
β_2 (RATIO018)	0.05	0.13	0.69
β_3 (RATIO027)	-0.03	0.01	0.03

Table 2: Results of estimation of SAR model