

A Hierarchical Spatial Model for Geographical Analysis of Business Mortality Risk

Un Modello Spaziale Gerarchico per l'Analisi Geografica del Rischio di Mortalità delle Imprese

Fabio Divino

Università di Roma "La Sapienza", Via del Castro Laurenziano, 9 – 00161 Roma,
fabio.divino@uniroma1.it

Maria Rita Sebastiani

Università di Roma "La Sapienza", Via del Castro Laurenziano, 9 – 00161 Roma,
mariarita.sebastiani@uniroma1.it

Riassunto: Lo scopo di questo lavoro è di proporre l'uso di un modello spaziale per l'analisi della variazione geografica del rischio di mortalità per le aziende.

Keywords: Business demography, Spatial model, Hierarchical model, MCMC.

1. Introduction

The aim of this paper is to present the use of a hierarchical spatial model to analyse the geographical variation of business mortality rates. Two main reasons underlie the importance of studying “how” and “why” the mortality of firms is differently distributed at a geographical level. First, the central economy authority of a country can be interested to know that in order to plan locally economic and fiscal politics (ex-ante analysis). Second, the same authority can be interested to control or to validate the effects of some legislative interventions (ex post analysis). Further, in the last years the European Union has got the use to use business demographic indicators (e.g. birth and mortality rates) to rank the risk of financial investments in the several countries of the Euro-area. This point is also related with the possibility for the European companies to access to different financial funds of EU. We will focus on the mortality of the Italian firms classified by size, age and economic field. When analysing data from a fine classification process, it can happen to have to handle sparse data matrices. If that is the case, it needs a model based approach in order to produce consistent statistical indicators. We propose a hierarchical spatial model and a Bayesian MCMC machinery to make inference on the geographical variation of the firm mortality incidence.

2. The business demography

Business demography studies the changes occurring within a population of firms, by considering the most important factors that can influence the demographic events. The mortality rate of a firm population is a good indicator on the health conditions of an economic system. The analysis of the firm mortality needs to take into account of some factors that can influence it; in Epidemiology such factors are named confounders. Many empirical studies show that the business mortality depends inversely on the firm

size. Another important factor is related to the production process, indeed the high capital intensive industries seem to have lower mortality rates than the high human intensive industries. There is also an age-period-cohort effect on the firm death incidence, from many empirical studies it results that the mortality rate generally decreases with respect to the firm age. Often, the firm mortality shows a geographical pattern: developed areas are opposed to areas more critical. The economists explain that by mean of the economy of agglomeration (facilities, natural resources, closeness to the productive factors, communication networks, lower costs) and by the endogenous regional growth. Our goal is to evaluate the geographical differences in mortality incidence, while controlling the confounding effect of age, size and economic activity.

3. A Hierarchical Spatial Model

When modelling geographical referenced data two sources of helpful information are expected to exist. First, information on spatial trend is often available by mean of covariates. A model based approach allows the estimation of the parameters associated with the explicative variables, e.g. regression coefficients. Second, adjacent area units are typically similar and homogeneous with respect to the features to be analysed. A spatial model uses this contextual information to modulate the global information (regression and spatial trend) in order to lead to more consistent estimates of the quantities of interest. We propose the spatial model introduced by Besag J., York J. and Mollie A. (1991) in the framework of image analysis and widely applied by other authors in disease mapping (see the book edited by Lawson et al.;1999), in ecological analysis (e.g. Clayton, Bernardinelli and Montomoli; 1993) and in biogeography (e.g. Penttinen, Divino and Riiali; 2003). Formally, if the study area A is divided into n subareas $A_1, \dots, A_n$; we denote by δ_i and Y_i respectively the firm mortality rate and the observable death count referred to the region A_i . Our hierarchical model consist of four levels: 1) the conditional likelihood model for the count Y_i given the mortality rate δ_i ; 2) the structural equation model for the rate δ_i including spatial trend, local homogeneity and size heterogeneity effect; 3) the prior distribution model for the regression parameters, fixed effects and random components; and 4) the hyperprior model for the parameters in the prior distributions. The first level considers the likelihood model. Conditional on the mortality rate δ_i , the observable count of closing firms is assumed to follow a Poisson distribution with mean given by $N_i\delta_i$, where N_i represents the known total number of the target firms in the region A_i during the observation time. By assuming conditional independence of the counts, the joint likelihood model is given by

$$l(y; \delta) \propto \exp \left\{ \sum_{i=1}^n (y_i \log \delta_i - \delta_i N_i) \right\}.$$

The second level considers the structural equation for the mortality rate δ_i . Following a spatial approach, we assume that each δ_i is smooth in the sense that mortality parameters of nearby areas do not differ drastically from each other. Further, observed covariates could be used to model the spatial trend, for example the geographical centroids of each region A_i . Third, an unstructured local variation may take place. We assume that all these components act on the logarithmic scale of the mortality rates in an additive way. Then the structural equation for the rate δ_i is given by

$$\log N_i \delta_i = t_i + u_i + v_i.$$

For specifying this equation, $t_i = \sum_k \beta_k z_{ik}$ is the fixed effect part containing the covariates effects. Second, $u = (u_i ; i=1, \dots, n)$ is assumed to be an intrinsic Gaussian random field defined by the probability density

$$p(u | \sigma_u) \propto \frac{1}{\sigma_u^n} \exp \left\{ -\frac{1}{2\sigma_u^2} \sum_{i \sim j} w_{ij} (u_i - u_j)^2 \right\},$$

where the sum in the exponential is over all pairs of neighbouring subareas (defined by the relation “ $\sim$ ”) and w_{ij} defines the spatial interaction weight for the pair of regions A_i and A_j . Finally the v_i for $i=1, \dots, n$ are unstructured (non-spatial) random effects which are assumed to be exchangeable and Normal distributed with mean zero and variance σ_v^2 . Their role is to model non-spatial extra variability due to the different firm population size in each region A_i . The above structural equation introduces a set $\beta_1, \dots, \beta_K$ of fixed effect parameters, we assume that they are independent and normal distributed with mean zero and variance $\sigma_1^2, \dots, \sigma_K^2$ respectively. The fourth level of the model considers the specification of the hyperpriors for the variance parameters $\sigma_1^2, \dots, \sigma_K^2$, σ_u^2 , and σ_v^2 . We assume they are Inverse Gamma distributed with proper fixed parameters. As a summary, we can derive by mean of the Bayes rule the posterior probability distribution $p(\beta_1, \dots, \beta_K, u_1, \dots, u_n, v_1, \dots, v_n, \sigma_1^2, \dots, \sigma_K^2, \sigma_u^2, \sigma_v^2 | y_1, \dots, y_n)$ in order to make inference on the quantity of interest, that are the rates δ_i for $i=1, \dots, n$. Due to the intractable nature of the above posterior probability, Bayesian inference on the firm mortality rates can be carried out by using a Markov chain Monte Carlo approach (Robert, C.P. and Casella G., 1999). The basic idea of the MCMC methods is to build a Markov chain whose stationary distribution is represented by the target posterior. Successively, the chain is simulated by the Monte Carlo sampling. After stationary is reached, the sequence of the drawn values can be considered as a dependent sample just from the posterior of interest and they can be used to make inference on the quantities of interest. In the construction of the Markov chain we propose a Gibbs-Metropolis sampler. Convergence has been controlled by the Gelman and Rubin statistics.

4. Data and results

We use our approach on real data referred to the small and medium Italian firms classified by size, age and economic field. In Italy, there are two important statistical sources for firm demography: the Business Register at the Unioncamere and the Osservatorio Statistico sulle Imprese at the INPS. We consider data from the INPS source in order to produce estimates of the business mortality rates at a provincial level during the late 1990s. Actually applications on real data are still in progress, but results from simulations suggest good fitting of the proposed approach. On a regular lattice of 50×50 pixels we simulated 4 types of spatially structured Poisson processes with the same geographical pattern but referred to different levels of the baseline mortality incidence: strongly low (0.01), low (0.05), medium (0.15) and high (0.3). They represent the true unknown models to be estimated. The choice of those baseline values has been done on the basis of some explorative analysis on real data. In fact, they can represent hypothetical death counts processes for different kinds of firm. For example

high baseline rate can be referred to start-up enterprises with few employed people. When the observed frequencies are low (low baseline rate and/or low population at risk) we expect an over-dispersed Poisson process. If this is the case standard estimates based only on the likelihood model show a noised behaviour. In the following Table 1 we present results from our simulations. We compare the maximum likelihood estimates (MLE) to the estimates from the proposed hierarchical Bayesian model (HBE). In columns 2 and 3 we present the global square error referred to the total of 2500 pixels. We carry also a measure of goodness of fit. For each kind of process we generated from the true unknown model further 10 simulations of counts. Then we measured the fit of ML model and of HB model by mean of the Chi square statistics. We crossed the predicted frequencies by each model with the simulated observations. In columns 3 and 4 are presented the means of Chi square referred to the 10 simulations of each model.

Table 1: *Results from simulations*

Baseline rate level	Global square error		Goodness of fit χ^2	
	MLE	HBE	MLE	HBE
Strongly Low (0.01)	0.25	0.19	177008403.7	2306382.3
Low (0.05)	1.21	0.44	769157.2	5345.1
Medium (0.15)	4.18	1.77	5810.8	4445.6
High (0.3)	7.57	3.58	5269.1	4205.1

From the Table 1 we can observe that the HB model performs, in term of global square error, ever better than ML model, even when the generating process is based on high baseline rate. In fact, if this is the case, we expect that MLE have to be quite consistent. It is important to note how bad are the performances of both the models when the true process is too over-dispersed. This is due to the fact that it is just enough to observe few cases (just one or two!) when the adopted model predict zero to have such bad fitting. This problem is under study by many researchers and common opinion is that models enclosing spatial components can be control and stabilise the estimates.

References

- Besag J., York J. and Mollie A. (1991) Bayesian image restoration with applications in spatial statistics, *Annals of the Institute of Mathematical Statistics*, 43, 1-21.
- Carrol G.R. and Hannan M.T. (2000) *The demography of corporations and industries*. Princeton University Press, Princeton, NJ.
- Clayton D.G., Bernardinelli L. and Montomoli C. (1993) Spatial correlation in ecological analysis. *International Journal of Epidemiology*, 22, 1193-1202.
- Contini B. and Revelli R. (1992) *Imprese, occupazione e retribuzioni al microscopio*. Il Mulino, Bologna.
- Evans D. (1987) The relationship between firm growth, size and age: estimates for 100 manufacturing industries, *Journal of Industrial Economics*, 35, 567-581.
- Penttinen A., Divino F. and Riiali A. (2003) Spatial small-area association of two species, In *Highly Structured Stochastic Systems*, Peter Green, Nils Hjort and Sylvia Richardson eds., Oxford University Press.
- Robert C.P. and Casella G. (1999) *Monte Carlo Statistical Methods*, Springer & Verlag, New York.