Location as risk factor Spatial effect in insurance

Ildiko Vitéz
Department of Probability Theory and Statistics
Eötvös Loránd University
Pázmány P. stny. 1/C Budapest, Hungary

Keywords: Spatial statistics. Markov random field. Markov Chain Monte Carlo (MCMC). Insurance
AMS: 60J80

Abstract

Our aim was to examine the territorial dependence of risk for household insurances. Besides the classical risk factors such as type of wall, type of building, etc., we consider the location associated to each contract. Three different methods are applied to describe the spatial effect: Markov Random Field model; a Potts model in a hierarchical Bayes-model; and a Maximum Likelihood estimation using Potts model. In the case of hierarchical Bayes-models we performed an iterative algorithm. As estimating all the effects jointly would result in enormous running time, instead of it we proceeded as follows: first fit a Generalized Linear Model (GLM) for the classical covariates, then fit the spatial model using Markov Chain Monte Carlo (MCMC) method. Iterate this procedure several times. We achieve much better fit by performing eight iterations.

1. Introduction - Risk models with spatial components

Analyzing an insurance data-set we assumed Poisson distribution for the number of claims of each client with the appropriate parameter. Estimating the values of these parameters we considered some classical variables such as type of wall, type of building, ... and as our purpose was the analysis of the spatial effect we involved the region associated to the contracts in our model. For the spatial effect we considered 3 models; Markov Random Field model; Potts model in a hierarchical Bayes model using some prior distributions for the parameters; and Potts model using Maximum Likelihood estimation for the spatial parameter-estimations. Basically there are two ways of dealing with the classical and the spatial risk factors; we can estimate the classical ones first, and then fixing their values as if they were known, estimate the spatial effect, or we can estimate all effects simultaneously. As the usual critic of the first method goes it is not reasonable to cope with the different variables separately, simultaneous estimation is more accurate. Though this latter concept seems to be more desirable but the realization can be hampered. If the model is several-level hierarchical Bayes model (eg.: the spatial effect is chosen to be random variable with some prior distribution) using MCMC algorithm to estimate the

parameters is often necessary. In this case too many parameters result in cumbersome full conditionals and enormous running time. To avoid this difficulty in the case of hierarchical Bayes model we suggest the following. Fit first a generalized linear model to the classical risk factors then keeping the received parameters fixed estimate the spatial effect; this way can work with lower dimensional vectors. Return then to the classical risk factors and refit the GLM keeping now the spatial effect fixed, and then move on to the spatial estimation with the newly fixed GLM parameters. By iterating these steps much better predictions can be obtained and the disadvantages of the distinct estimation of the classical and the spatial risk factors can largely be eliminated. In the case of Maximum likelihood estimation we use the effects of the classical variables obtained from the GLM. As result we receive untractable expressions so we apply Expectation Maximization (EM) algorithm. For the calculations we use Monte Carlo EM algorithm. The paper is organized as follows. Section 2 describes the basic model. In Section 3, 4 and 5 we describe the models applied to the spatial effects.

2. Model construction

Selecting the explanatory variables in the GLM we found that: "time spent in risk", "type of building" (4 types), "type of wall" (4 types), "type of roof" (6 types), "type of tarif" (4 categories), and "population size of the locality" (10 groups) are the significant classical risk factors. As our aim is to analyze the effect of location we include the effect of the regions (168) into our model. At the lowest level of the hierarchy we suppose the number of claims to comply with the Poisson distribution:

$$y_i \sim \ Poisson(t_i \cdot E_i \cdot \lambda_{r_i})$$

where y_i denotes the observed number of claims of client i during the insured period, $i = 1 \dots n$, n is the number of contracts, E_i is the expected number of claims per day based on the classical variables, t_i is the number of exposure days i.e. the days spent in risk, $\lambda_k = exp(u_k)$, where u_k is the relative risk of region k, and r_i denotes the index of the region that contract i belongs to. We used a generalized linear model for approximating the effect of the classical risk factors. Using the received parameters we calculated the E_i values of each client:

$$E_i = exp(\beta_0 + \beta_{b_i}^{building} + \beta_{w_i}^{wall} + \beta_{t_i}^{tarif} + \beta_{r_i}^{roof} + \beta_{p_i}^{pop})$$

[1] As for the spatial effect we chose to work with correlated regions. The reason is that if no claims incur or no data of claims is available in a region using

[1] where $\beta^{building}$ is the vector of the effects of the different types of building, and b_i denotes the type of building belonging to client i. The other notations can be interpreted similarly.

uncorrelated spatial effects the estimated relative risk for that region would be zero, which is clearly not the best estimate. So working with correlated regions has the advantage that neighbouring regions influence the relative risk of each other - which is a sensible assumption. To create a neighbourhood-system we regard two regions as neighbours if the distance between their centers is less than 35 kms. The choice of the distance ensures that all regions have at least one neighbour.

3. Markov Random Field model

In our first model the prior distribution of the relative risks of the regions - denoted by $\{u_i, i = 1, 2, \cdots, m\}$, $m = 168$ being the number of regions - is given by a Markov Random Field model. In the Markov Random Field model the relative risk of region i is *normally* distributed with the averaged relative risk of the neighbouring regions as expected value and with variance which is an overall variance parameter σ^2 divided by the number of neighbours:

$$\left[u_i | u_{j,j\neq i}, \sigma^2\right] \sim N\left(\frac{1}{|\delta_i|}\sum_{j\in\delta_i} u_j, \frac{\sigma^2}{|\delta_i|}\right)$$

where the hyperparameter σ - denoting the strength of correlation - is a random variable with *inverse* $Gamma(\alpha, \lambda)$ prior distribution, δ_i is the set of the neighbours of region i.

Using Bayes theorem we determine the posterior distributions of the parameters. For the relative risk of the regions we receive the following:

$$\left[u_i | u_{j,j\neq i}, \sigma^2, \underline{y}, \underline{E}\right] \sim \prod_{k=1, r_k=i}^{n} exp\left(-t_k E_k exp(u_i) + y_k u_i - \frac{|\delta_i|(u_i - \bar{u}_i)}{2\sigma^2}\right)$$

where $\bar{u}_i$ denotes $\frac{1}{|\delta_i|}\sum_{j\in\delta_i} u_j$.
The posterior distribution of σ^{-2} is: $Gamma(\alpha + \frac{168}{2}, \lambda + \frac{1}{2}\underline{u}' K \underline{u})$ where K is a 168×168-dimension matrix, with elements: $k_{r,r} = \mid \delta_r \mid$ in the main diagonal, and $k_{r,s} = -1$ if region r and s are adjacent, otherwise $k_{r,s} = 0$. Clearly the formula obtained for the posterior distribution of the relative risk is not the density function of a well-known distribution so we use MCMC simulation to get the expected values. Once we have the spatial effects we reestimate the effects of the classical factors fitting GLM, and then refit the spatial model, etc.. Estimating the spatial effect after dealing with the classical risk factors enables us to lower the dimension of vectors used in the MCMC simulation. Denoting the logarithm of the Poisson parameter of client i by ν_i we get the following expression:

$$\nu_i = (log(t_i) + \beta_0 + \beta_{b_i}^{building} + \beta_{w_i}^{wall} + \beta_{t_i}^{tarif} + \beta_{r_i}^{roof} + \beta_{p_i}^{pop}) + u_{r_i}.$$

Grouping the non-spatial parameters and the time together we can regard the expression $t'_i = t_i \cdot exp(\beta_0 + \beta_{b_i}^{building} + \beta_{w_i}^{wall} + \beta_{t_i}^{tarif} + \beta_{r_i}^{roof} + \beta_{p_i}^{pop})$ as a modified variable of time. Using this variable we get a simpler expression for ν_i:

$$\nu_i = log(t'_i) + u_{r_i}.$$

In this step distinguishing the contracts just by the regions we can lower the dimension. Let's summarize the modified version of times for each region and also the number of claims for each region. This way we can work with these vectors of only 168 elements: t_i^r, y_i^r. For determining the expected value of vector u we generate random samples from its density function and get an estimate by their mean. In our process we iterated these two steps. Returning to the parameters β we get new estimates for them taking the spatial effect into consideration. We refit the Generalized Linear Model the same way as we described before, modifying only the vector of times spent in risk with the effect of regions: $t'_i = t_i \cdot exp(u_{r_i})$. Eigth steps of iteration seemes to be enough to get good estimates.

4. Potts model in a hierarchical Bayes model

In this model for the relative risk of the regions we assumed Potts model [4]. $u_i = u_{z_i}$, where u_i characterize k different components, and z_i are allocation variables taking values in $1, 2, ..., k$. Let's denote the set of all possible allocations by Z; $Z = \{1, 2, \ldots, k\}^n$

In the Potts model formulation, the z_i are modeled jointly,

$$p(z|\psi) = e^{\psi \cdot U(z) - \theta_k(\psi)},\ U(z) = \sum_{i \sim i'} I[z_i = z_{i'}],\ \theta_k(\psi) = log(\sum_{z \in Z} e^{\psi U(z)}).$$

The prior distributions of parameters λ are Gamma, and that of parameter ψ is discrete uniform. Further details of the model can be found in [4].

5. Maximum Likelihood estimation using Potts model

Assuming Poisson distribution for the number of claims of a certain client, using Potts model for the spatial effect with the law of total probability for the values of z we deduce the distribution of y. Performing a Maximum Likelihood estimation for the parameters λ and σ we received the following expressions:

$$\sum_{z \in Z} c_h(z) \cdot \lambda_h^{\nu_h(z)-1} \cdot e^{-\lambda_h \cdot \epsilon_h(z)} \cdot (\nu_h(z) - \epsilon_h(z) \cdot \lambda_h) = 0$$

for λ, where

$$\nu_h(z) = \sum_{z_i=h} y_i,\ \epsilon_h(z) = \sum_{z_i=h} t'_i,\ c_h(z) = e^{\psi \cdot U(z)} \cdot \prod_{j \neq h} e^{-\lambda_j \cdot \epsilon_j(z)} \cdot \lambda_j^{\nu_j(z)}$$

and for $x = e^{\psi}$ we need the maximum of the following expression:

$$\frac{\sum_{z \in Z} x^{U(z)} \cdot \prod_{i=1}^{n} e^{-\lambda_{z_i} \cdot E_i} \cdot \lambda_{z_i}^{y_i}}{\sum_{z \in Z} x^{U(z)}}.$$

In both cases we received an implicit formula for the parameters. To obtain more tractable expressions we applied EM algorithm. In each step we receive new values for the parameters using the following equations:

$$\lambda_j = \frac{\sum_{z \in Z} P(z|y, \theta_t) \cdot \nu_j(z)}{\sum_{z \in Z} P(z|y, \theta_t) \cdot \epsilon_j(z)}$$

and for $x = e^{\psi}$

$$\sum_{z \in Z} x^{U(z)} \cdot \{U(z) - \sum_{z' \in Z} U(z) \cdot P(z'|y, \theta_t)\} = 0,$$

where

$$P(z|y, \theta_t) = \frac{e^{\psi \cdot U(z)} \cdot \prod_{i=1}^{n} e^{-\lambda_{z_i} \cdot E_i} \cdot \lambda_{z_i}^{y_i}}{\sum_{z \in Z} e^{\psi \cdot U(z)} \cdot \prod_{i=1}^{n} e^{-\lambda_{z_i} \cdot E_i} \cdot \lambda_{z_i}^{y_i}}.$$

As in all summations the number of terms is exponential working with high dimensional parameters we use Monte Carlo Expectation Maximization (MCEM) algorithm. In each EM step we work with an estimate of the function to be maximized [2]. The results of the different models applied on an insurance data set will be presented on EYSM 2007.

6. Bibliography

[1] Arató, N. M., I. L. Dryden, C.C. Taylor (2004). *Hierarchical Bayesian modelling of spatial age-dependent mortality.*

[2] Caffo, B., S., Jank, W., Jones G. L. (2004). *Ascent-Based Monte Carlo EM.*

[3] Dimakos, X.K., Frigessi di Rattalma, A., (2002). *Bayesian premium rating with latent structure.*, Scandinavian Actuarial Journal, 162-184.

[4] Green, P.J., Richardson, S., (2002). *Hidden Markov Models and Disease Mapping.*, Journal of the American Statistical Association, 94., 460.