SPATIAL DISCRETE-TIME EVENT HISTORY MODELS: AN APPLICATION TO HOME-LEAVING

Riccardo Borgoni and Francesco C. Billari

Max Planck Institute for Demographic Research,
Doberaner Str. 114, D-18057 Rostock, Germany
e-mail: {borgoni,billari}@demogr.mpg.de

ABSTRACT. In this paper we deal with the use of computationally complex methods in the spatial analysis of discrete-time event histories. We present an application to home-leaving data of young Italian adults. The main goal is to explore the spatially-related patterns of behaviour at a relatively fine level, by using a flexible model in conjunction with MCMC inference. We also build maps of the pattern.

1 INTRODUCTION

Young Italian adults leave their parents' home at rather different ages. The median age at leaving home, at a broader regional level, has great variation (Billari and Ongaro (1999)). Such differences are usually explained by referring to historical preferences on the living arrangements, and to economic and institutional factors.

Two main issues have not been dealt so far in the home-leaving literature. First, much less is known at a finer spatial level, and a complete map of the geographical pattern is still to be built. Second, an analysis which exploits recent opportunities for the study of spatially-located data has not yet been done.

This paper is concerned with the use of computationally complex methods in the spatial analysis of discrete-time event histories as applied to data on leaving home of young Italian adults. The paper is structured as follows. In section 1, we introduce the data set we use. Then, we briefly describe the methods (GAMM and the MCMC procedures). Our results are presented in section 4. Section 5 concludes the paper giving also further perspectives.

2 DATA

The data we use come from self-reported retrospective interviews of the Italian Fertility and Family Survey (De Sandre et al. (1997)), part of a program implemented by the Population Activities Unit of the United Nations Economic Commission for Europe. A representative sample of 6,030 residents (4,824 females and 1,206 males) born between 1946 and 1975, were interviewed about the timing of life course events. Each respondent was asked on whether he/she had ever left the parental home to start living on his/her own. If yes, he/she was asked the date of such event.

As the main goal of this analysis is to grasp the spatial pattern of the phenomenon, if there is any as we expect, we link the individual-level data set with a data set containing latitude and longitude of the centroids of each italian province. For each individual we use the spatial

location of the main residence within his/her first 15 years of age. We drop from our analysis all those respondents who either did not answer on whether the event has happened, or did not report the time of the event, or whose place of residence was unknown or outside Italy. This reduces the data set to 5862 observations.

In order to use a discrete-time event history model, data must be arranged in a suitable way. Thus we develop a person-year data set, where all subjects are enter the period at risk at the age of 14, and where an indicator variable for the event at age t is introduced. As the event under study is an absorbing event, respondents either disappear from the data set after having experienced it, or they are censored at the age corresponding to the interview. As we have considered retrospective data, we do not have items lost to follow up, and the only type of censoring is right truncation. With our data, we obtained a person-year file of 63663 records.

3 METHODS

Generalised additive models (GAM) (Hastie and Tibshirani (1990)) extend generalised linear models (McCullagh and Nelder (1989)) by replacing the linear predictor with an additive one, η, allowing for non linear effects. In these models, the conditional expected value, μ, of response variable, Y, is linked to η through a regular link function $g(\mu) = \eta$.

Extending these models to longitudinal data in discrete time is relatively straightforward. In a conditional framework, past observations of the response, as well as of any time-varying explanatory variable, can be introduced in η. More recently generalised additive marginal models have been developed (see Fahrmeir and Tutz (2001) for a full list of bibliographic references).

Sampling units according to some spatial features clearly induces a form of hierarchy. Spatial clustering is a well-known problem in many different fields as in epidemiology (Wakefield et al. (2000)), labor market (Fahrmeir and Lang (2001-B)) or demography (Steele et al. (1996)).

To take into account the intra-group variability, we can introduce one or more random effects. Cluster-specific effects are assumed to be i.i.d. according to a mixing distribution. This approach can be extended to more than one nested (or even cross-classified) levels. This so-called multilevel approach is fully described, among others, in Goldstein (1995).

Introducing random effects in a GAM defines a generalised additive mixed model (GAMM).

Sometimes, the predictors can be split up in a set of q metrical or spatial covariates and a vector of s further covariates W, leading to the semi-parametric form:

$$g(\mu) = \sum_{j=1}^{q} f_j(X_j) + W'\gamma + U_g \qquad g = 1, \ldots, G$$

where $U = \{U_1, \ldots, U_G\}$ is a group-specific random effect.

Actually, the i.i.d. assumption is not always appropriate when we deal with geographical clustering, as the unobserved heterogeneity is often caused by behavioral factors that tend to be similar among close areas.

A common way to deal with spatial covariates is to assume that neighbouring areas (i.e. areas sharing a boundary) are more alike than others. Specifically, we refer to the case

where area-specific random effects, $B = \{B_1, \ldots, B_G\}$, are spatially correlated. We assume that $B_s|\{B_r : r \neq s, \tau\} \sim N(\sum_{(r\in\partial_s)} w_{rs}/w_s B_r, \tau^2 w_s)$ where ∂_s is the set of the neighbours of s and $w_s = \sum_{(r\in\partial_s)} w_{rs}$. This model is sometime referred to as an *auto-normal* or as an *intrinsic Gaussian conditional autoregressive* (CAR) model (Besag et al. (1991)). We assume $w_{rs} = 1$ if areas r and s are adjacent and $w_{rs} = 0$ otherwise (and $w_{ss} = 0$).

This kind of models can be hardly dealt with in a standard likelihood framework. On the contrary, a bayesian approach allows a natural way to their estimation. In such a context, γ and $f_1, \ldots, f_q$ are supposed to be random variables, and prior distributions must be specified for them, as well as for random parameters and their hyperparameters. Fahrmeir and Lang (2001-A) give a broad review of the literature this topic.

In the case study presented in this paper, we model the hazard of leaving the parental home as a logit model:

$$\lambda(t|\eta_{it}) = P(Y_{it} = 1|\eta_{it}) = \exp(\eta_{it})/(1+\exp(\eta_{it})).$$

The covariates are both time-varying, as age (T), and time-constant, as gender (SEX), province of residence and cohort (COH, binary: born before or after 1960) of the respondent. As space is considered as a proxy of unobserved influential factors, some of them acting locally and others obeying to a stronger spatial structure, it is useful to specify both structured and unstructured spatial effects in a model. Spatial information are present in the model as a structured (B), and an unstructured (U) random effect, as well as metrical covariates trough the latitude (LT) an longitude (LG) of the province centroids. Thus the model is specified, for every item i, as:

$$\eta_{itg} = \alpha_0 + T_i + \alpha_2 SEX_i + \alpha_3 COH_i + f_1(LT_{ig}) + f_2(LG_{ig}) + U_{ig} + B_{ig}.$$

where g is the province index ($g = 1, \ldots, 95$), B follows a CAR process and U is the unstructured random effect assumed to be gaussian 0 mean distributed with variance σ^2. Finally f_1 and f_2 are bayesian P-splines (Lang and Brezger (2000-A)) of first degree, with equally spaced knots starting from the southern- and from the western-most centroids respectively.

As the number and positions of knots in a spline smoother can strongly affect the amount of smoothing, the idea of P-splines (Ellers and Marx (1996)) is to start with a relatively large number of knots, preventing overfitting by penalising flexibility through finite differences of coefficients of adjacent B-splines. Bayesian P-splines consists in replacing finite differences with their stochastic counterpart, i.e. random walk processes of the same order.

We assume a diffuse prior for the first parameter of the random walk processes and a normal distribution for innovations.

The model set up is completed by assuming an $IG(a,c)$ law for the precision parameter τ of the CAR model, and the independence of observations given the model parameters and among priors.

Since the posterior distribution is often numerically intractable, the MCMC approach consists in creating a Markov chain whose iterations converge to this distribution. Sampled values are used to estimate the characteristics of the posterior. The main idea consists in factorising the posterior in several terms using the model assumptions and sampling each full conditional given the rest and the data.The sampler we used is fully described in Farhmeir and Lang (2001-A). For updating the values of the smoothing functions, a very efficient

Metropolis-Hastings algorithm is used, where blocks of parameters are update instead of one. For fixed and unstructured random effects a slightly modified version of Gamerman's weighted last square proposal (Gamerman, 1997) is used.

At first, we run a quite long MCMC experiment (50000 iterations) to explore the correlations among sampled values and small experiments to chose the block size. A final Monte Carlo experiment is run with 32000 iterations (2000 for the burn-in period). A graphical inspection of the sampled values shows a very good convergence of the chain.

All analyses presented in this paper are performed using *BayesX* (Lang and Brezger (2000-B)).

4 RESULTS

The results of our main analysis are presented in Tables 1. The reference classes for categorical variables are: age 14-15, female and older cohort. The log-odds of home-leaving increase with age (other than the last category), while they are significantly lower for males and members of the younger cohort. Figure 1 shows the spatial effects due to geographical

Variable	mean	Std. Dev.	10 % quant.	median	90% quant.
const	-3.91	0.088	-4.03	-3.91	-3.80
AGE: 16-17	0.31	0.112	0.17	0.31	0.45
AGE: 18-19	1.57	0.093	1.45	1.57	1.69
AGE: 20-21	2.04	0.090	1.93	2.04	2.16
AGE: 22-23	2.28	0.091	2.16	2.27	2.40
AGE: 24-25	2.56	0.094	2.44	2.56	2.68
AGE: $\geq$ 26	1.95	0.091	1.84	1.95	2.07
SEX (male)	-0.42	0.046	-0.48	-0.42	-0.36
COHORT (young)	-0.48	0.038	-0.53	-0.48	-0.43

Table 1. Fixed effects

coordinates with their posterior 10-th and 90-th quantiles . It is rather evident that moving from the South to the North of Italy (latitude), the log-odds decrease almost monotonically. A West-East (longitude) effect is also evident.

In Figure 2, maps depict structured and unstructured spatial effects. We can notice that, 1) no trend is present in spatial effects and, 2) almost all random effects (both structured and unstructured) are not significantly different from zero, with some exceptions of trend-breaking. The estimated variance parameter for the unstructured random component is 0.0223, while the estimated variance of the structured one is 0.01. Thus, a higher share of spatial variability left is due to the unstructured heterogeneity.

5 CONCLUSION AND FURTHER PERSPECTIVE

In this paper we use computationally complex methods in the spatial analysis of discrete-time event histories. We show an application to data on leaving home of young Italian adults. As

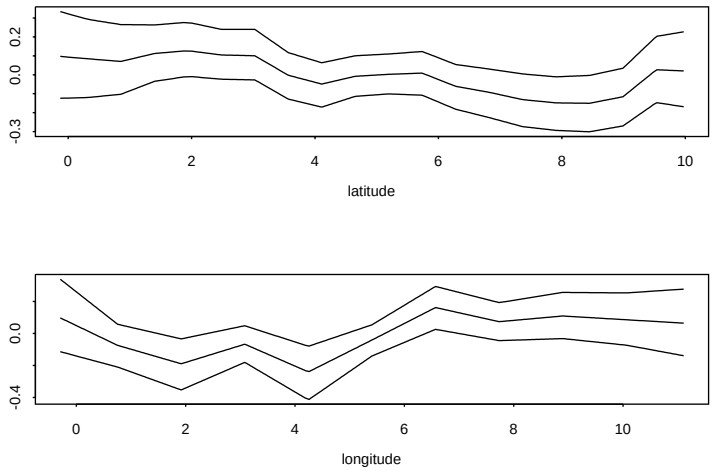

Figure 1. Spatial trend effects with 80 % credible regions. The origin of coordinates is shifted to the minimum value of province centroids

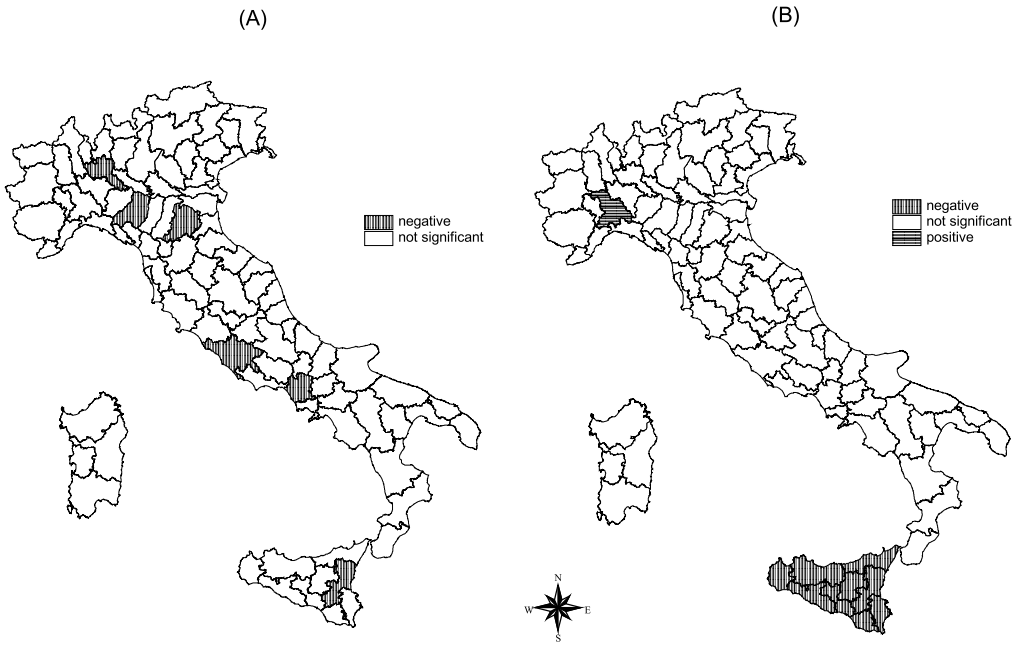

Figure 2. Random effects: (A) unstructured, (B) structured

more detailed spatial information are available from the survey used in this paper, further steps shall concern the analysis to more local pattern (say at municipality levels), as well as some simulation based investigations to assess the performance of the proposed models in such a framework.

REFERENCES

BILLARI, F.C., ONGARO, F. (1999): Lasciare la Famiglia di Origine: quando e perché? In: P. De Sandre, A. Pinnelli, A. Santini: (Eds.) *Nuzialità e Fecondità in Trasformazione: Percorsi e Fattori del Cambiamento*, il Mulino, Bologna, 327-346.

DE SANDRE, P., ONGARO, F., RETTAROLI, R., SALVINI, S. (1997): *Matrimonio e figli: tra rinvio e rinuncia*, il Mulino, Bologna.

ELLERS, P.H.C., MARX, B.D., (1996): Flexible smoothing with B-splines and penalties, *Statistical Science, 2*, 89-121

GAMERMAN, D. (1997): Efficient sampling from the posterior distribution in generalised linear model, *Statistics and Computing*, *7*, 57-68.

FAHRMEIR, L., LANG, S. (2001) (A): Bayesian inference for generalised additive mixed models based on Markov random field priors, *Applied Statistics*, *50*, 201-220.

FAHRMEIR, L., LANG, S. (2001) (B): Bayesian semiparametric regression analysis of multicategorical time-space data, *Annals of the Instute of Statistical Mathematics*, *53*, 10-30.

FAHRMEIR, L., TUTZ, G. (2001): Multivariate Statistical Modelling Based on Generalised Linear Models, Springer, Berlin.

GOLDSTEIN, H. (1995): *Multilevel Statistical Models*, Second Edition, Edward Arnold, London.

HASTIE, T.J., THIBSHIRANI, R.J. (1990): *Generalised Additive Models*, Chapman & Hall, London.

LANG, S., BREZGER, A. (2000) (A): Bayesian P-splines, *Proc.of the 15th International Workshop of Statistical Modelling*, Bilbao.

LANG, S., BREZGER, A. (2000) (B): *BayesX*, University of Munich, Munich.

MCCULLAGH, P., NELDER, J.A. (1989): *Generalised Linear Models*, (2nd ed.) Chapman & Hall, London.

STEELE, F., DIAMOND, I., AMIN, S. (1996): Immunization uptake in rural Bangladesh: a multivel analysis, *Journal of the Royal Statistical Society*, *159*, 289-299.

WAKEFIELD, J.C., BEST, N.G., WALLER, L. (2000): Bayesian Approach to Disease Mapping. In: P. Elliott, J. C. Wakefield, N.G. Best and D. J. Brings: (Eds.) *Spatial Epidemiology: Methods and Applications*. Oxford University Press, Oxford, 104-127.