SPATIAL2

Spatial Data Methods for Environmental and Ecological Processes - 2nd Edition

PROCEEDINGS

EDITOR: Barbara Cafarelli

ENVIRONMETRICS

A Spatio-temporal model for cancer incidence data with zero-inflation [1]

Monica Musio
Department of Mathematics, University of Cagliari, Italy, mmusio@unica.it

Erik A. Sauleau
Faculty of Medicine, University of Strasbourg, France

Abstract: In this work we consider a joint space-time model for cancer incidence, using data on prostate cancer collected between 1988 and 2005 in a specific area of France. Our aim is to take into account possible non linear effects of some covariates and zero-inflation due to data aggregation for Poisson regression. We assume that counts of cancer cases follow zero-inflated Poisson distribution, where the probability of zero inflation is a monotonic function of the mean. The purpose of our analysis is to check whether the French prostate screening programme, which begins in 1994, results in a spatial or a spatial-temporal change of the pattern of the disease.

Keywords: Spatio-temporal model, cancer incidence data, zero inflation

1 Introduction

Cancer registries represent epidemiological instruments which are aimed at providing population based cancer incidence and mortality summaries. Usually the data are stratified by age group, year and geographical unit of residence. As the counts of cancer cases are distributed according to these variables, the dataset exhibits a proportion of zeros higher than would be expected under the Poisson distribution. The problem is also known as zero-inflation (Lachenbruch, 2002) and is common in ecological studies. We make the assumption, justified by the nature of the data analyzed, that the probability of zero inflation depends on the set of stratified variables. In this work we analyse data on prostate cancer incidence collected between 1988 and 2005 in the North-East of France. We present an approach to analyze the space-time evolution of the disease taking into account also possible non linear effects of other covariates (such as age) and the zero inflation due to extra Poisson variation. Prostate is a type of cancer which usually does not have a spatial distribution. Here we are interested in the space-time evolution of the disease to investigate if the prostate screening programme started progressively in the region since 1994 has a direct implication on the space or space-time evolution of the cancer.

[1]The second author was partially supported by *Visiting Professor program* of "Regione Autonoma della Sardegna"

2 Materials and Methods

Our data consists of all cases of prostate cancer (C61.- in the ICD-10 classification) diagnosed between the 1st January 1988 and 31st December 2005, in the region of Haut-Rhin in France. The total number of cases is 6878. The distribution of the number of cases aggregated over age groups (9 categories), across the 26 geographical units, each year has mean of 14.2 cases while the median is 10. Due to covariates, the data set counts were spread over 4374 cells with 1935 zeros (44% of the cells are equal to zero). Our objective is to detect effects of time, space, age and age-time interaction on the number of new prostate cancer cases, taking into account an high proportion of zero counts. We thus build different zero-inflated models and compare them using marginal likelihood.

Zero-inflated Poisson data are often analyzed via a mixture model specifying that the response variable, Y, comes from a mixture of 0 with probability ω and a regular Poisson component of mean λ with probability $1-\omega$ (Lambert, 1992).

Covariates may then enter into the model through the mean λ and/or through the probability ω. Here we consider a zero-inflated generalized additive model (Chiogna and Gaetan, 2007), where the mean of the regular component and the probability of zero-inflation are each modeled as a function of some nonparametric smooth predictors. As usual we assume that the mean of the Poisson distribution λ is equal to $\mathrm{E}(\mu)$ where E indicates expected number of cases under direct standardization and μ is the relative risk. For the log risk we consider the following linear predictor:

$$\log(\mu_{atr}) = \eta_{atr} = f_1(age_a) + f_2(year_t) + f_3(age_a, year_t) + f_4(east_r, north_r) \quad (1)$$

$a \in \{1,\ldots,9\}$, $t \in \{1,\ldots,18\}$, $r \in \{1,\ldots,26\}$, $f_1(\cdot)$, $f_2(\cdot)$ are smooth functions of the covariates age and year modeled using cubic regression splines, $f_4(east_r, north_r)$ is a thin plate regression spline, while, for modelling the smoothed age-time interaction, we use tensor products allowing smoothness parameter selection to be independent of the different scale of the covariates (for more details see (Wood, 2006)). We make the assumption that the probability of zero inflation is a linear function of the covariates. We are in the framework of constrained zero-inflated generalized additive model (COZIGAM) ((Liu and Chan, 2010)). In particular we consider the following two specifications:

1. Model 1: the dependence is constrained in such a way that the probability of zero inflation is linearly related to linear predictor. We have:

$$logit(\omega_{atr}) = \alpha + \delta\eta_{atr};$$

2. Model 2: the proportional constraint can be generalized by assuming that the proportionality constant is specific to each additive component, specifically:

$$logit(\omega_{atr}) = \beta + \delta_1 f_1(age_a) + \delta_2 f_2(year_t) + \delta_3 f_3(age_a, year_t) + \delta_4 f_4(east_r, north_r).$$

In both model the linear predictor is specified as in equation (1).

Because there is no closed form for the marginal likelihood, Laplace method is used to approximately compute the likelihood (Liu and Chan, 2010). The analyses have been performed using the R package `COZIGAM` (Liu and Chan, 2010), relying on `mgcv` package (Wood, 2001).

3 Results

According to the marginal likelihood, the best model is Model 2. In Table 1 are reported the values of the significant proportionality coefficients estimates for the best model, which provides strong evidence of a significant relationship between these smooth components in the mean of the non-zero-inflated distribution and in the zero-inflation probability, on their link scales. These values emphasize the main role that age plays on the zero-inflation, compared with the effect of time.

Figure 3 displays the smooth function estimates of Model 2. We can see that:

- The estimate of the time effect shows an increase of incidence up to 1995 then a strong decrease up to 2001 then an increase.
- the combined effect of age and time is quite relevant, in particular a progressive decrease in the age for the maximum incidence along time is evident.
- The estimated spatial effect is slightly significant. Except some boundary effects, there is a little peak of incidence in the north of the region (where a city of around 70,000 inhabitants is) and again a peak on the south-east part, difficult to separate from the boundary effect. Adding the spatio-temporal interaction in the model mod4 yields a non-significant effect.

	Covariate	estimate	standard error	pvalue
β		3.9939	0.85714	$p < 0.00001$
δ_1	s(age)	0.8859	0.04692	$p < 0.00001$
δ_2	s(year)	2.595	1.150	$p = 0.024$
δ_3	s(age, year)	1.237	0.29	$p < 0.00001$

Table 1: Significant coefficient estimates of the constrained generalized additive model

4 Concluding remarks

Zero-inflated generalized additive model provides a method for modeling incidence data by taking simultaneously into account possible non linear effects of continuous covariates and the spatio-temporal evolution of the disease. The number of extra zeros seems in particular linked to the age group. The aim of such study is to check

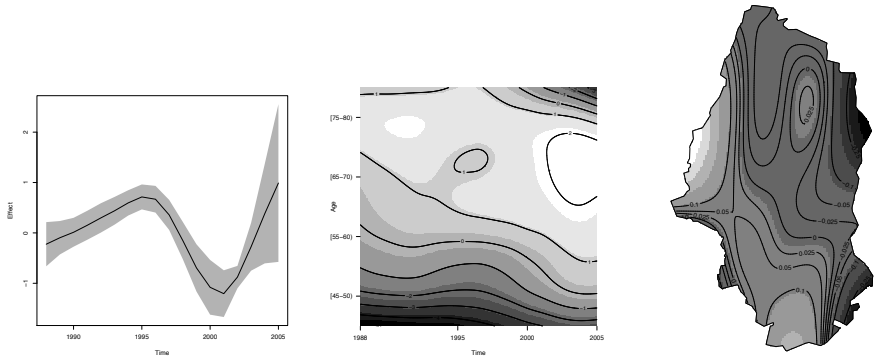

Figure 1: Effect of time, joint effect of age and time and spatial effect estimated for Model 2

whether the spatial pattern of incidences changes over time. The main finding is that there is a strong temporal effect, while the spatial effect is not very strong (not quite significant) and the spatial effect does not change over time (the space-time interaction was not significant). If we link the aspect of the main temporal effect with the development of the screening, it seems that the effect on the prostate cancer incidence is relevant since 1998 whereas the beginning of the organized screening campaign is 1994. This difference is probably due to a certain time for the screening programm to be fully efficient in the population.

References

Chiogna, M. and Gaetan, C. (2007). Semiparametric zero-inflated poisson models with application to animal abundance studies. *Environmetrics*, 18:303–314.

Lachenbruch, P. (2002). Analysis of data with excess zeros. *Statistical Methods in Medical Research*, 11:297–302.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14.

Liu, H. and Chan, K. (2010). Introducing COZIGAM: an R package for unconstrained and constrained zero-inflated generalized additive model analysis. *Journal of Statistical Software*, 35(11):1–26.

Wood, S. (2001). mgcv: GAMs and generalized ridge regression for R. *R News*, 1(2):20–5.

Wood, S. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62:1025–1036.

Then we can estimate θ by $\widehat{\theta}_S$, the root of the *estimating equation*

$$s(x, \theta) = 0. \tag{1}$$

When S is the log score, this is just the likelihood equation, and $\widehat{\theta}_S$ is the maximum likelihood estimate. More generally, for any differentiable scoring rule and any smooth statistical model, $\mathrm{E}_\theta\{s(X, \theta)\} = 0$, *i.e.* (1) is an unbiased estimating equation (Dawid and Lauritzen 2005). In particular it will typically deliver a consistent, if not necessarily efficient, estimator in repeated sampling. We can then choose S to increase robustness or ease of computation.

In the context of a spatial process $X = (X_v : v \in V)$, we can define a useful class of proper scoring rules (Dawid *et al.* 2011) by

$$S(x, Q) = \sum_v S_0(x_v, Q_v), \tag{2}$$

where Q_v is the conditional distribution of X_v, given the values $x_{\setminus v}$ for the variables $X_{\setminus v}$ at all sites other than V, and S_0 is a proper scoring rule for the state at a single site. In particular, if Q is Markov on a graph $\mathcal{G}$, then Q_v only depends on the values $x_{\mathrm{ne}(v)}$ at the sites neighbouring v. This avoids the need to evaluate the normalising constant of the full joint distribution Q.

Corresponding to (2) we have estimating equation

$$\sum_v s_0(x_v, P_{\theta,v}) = 0 \tag{3}$$

with each term in the sum having expectation 0. When S_0 is the log score, (3) gives the (negative log) *pseudo-likelihood* (Besag 1975). For X_v binary and S_0 the quadratic ("Brier") score, it yields the method of *ratio matching* (Hyvärinen 2007).

Missing data are readily dealt with (although with some loss of efficiency). Let $A_v = 1$ if any value in $\{v\} \cup \mathrm{ne}(v)$ is missing. Then so long as the data are missing completely at random, $s_0(x_v, P_{\theta,v}) \times A_v$ has expectation 0, so we can just omit incomplete terms from (3) while retaining an unbiased estimating equation.

3 Phytophthora data

Figure 1 displays the presence or absence of the pathogen *Phytophthora capsici* Leonian in bell pepper plants on a regular 20×20 grid (Chadoeuf *et al.* 1992).

We model the data as a stationary first-order Markov process with respect to the grid, which thus follows the autologistic model (Besag 1972; Besag 1974; Gumpertz *et al.* 1997):

$$\operatorname{logit} \pi_{ij} = \alpha + \beta(x_{i-1,j} + x_{i+1,j}) + \gamma(x_{i,j-1} + x_{i,j+1}) \tag{4}$$

where π_{ij} is the probability of $X_{ij} = 1$, given all other values.

SPATIAL2

ISBN 978-88-96025-12-3

CDP SERVICE EDIZIONI