

Modeling issues in bivariate disease mapping ^(*)

La mappatura congiunta del rischio di malattia di due popolazioni

E. Dreassi¹, D. Catelan^{1,2}, C. Lagazio², A. Biggeri^{1,3}

¹ Dipartimento di Statistica “G. Parenti”, Università di Firenze
e-mail: dreassi@ds.unifi.it

² Dipartimento di Scienze Statistiche, Università di Udine

³ UO Biostatistica, CSPO Istituto scientifico regionale, Firenze

Riassunto: Analizzare congiuntamente la distribuzione spazio-temporale di una malattia in due gruppi disgiunti di popolazione permette di fare inferenza su eventuali andamenti comuni ed effettuare ipotesi sulla condivisione di fattori eziologici. In questo lavoro si analizza la mortalità per tumore al polmone nei Comuni della Toscana per gli uomini e le donne appartenenti alle coorti di nati dal 1905 al 1940. Si utilizza una serie di modelli Bayesiani gerarchici spaziotemporali con componenti di varianza comuni e specifiche per genere.

Keywords: Space-time hierarchical Bayesian models, bivariate disease mapping.

1. Introduction

Bivariate disease mapping is defined as the joint modeling of the spatial or spatio-temporal occurrence of two diseases or of the same disease in two subsets of the population at risk. It is of interest when inference aims to clarify the existence of a common feature and what is the extent of any specific variation.

When analyzing bivariate data the natural history of the disease(s) has to be taken into account in order to avoid invalid conclusions from exploratory exercises on aggregate data.

We focus on mortality for lung cancer by municipality and birth cohort. Several papers have been published on this topic, describing the progression of the epidemic from the end of nineteenth century up to nowadays (see for Tuscany Region, Lagazio *et al.*, 2003). The spatio-temporal pattern of lung cancer occurrence reflected the prevalence of occupational and life-style factors (e.g. smoking habits) with striking differences between males and females. Among females the role of occupational carcinogens is less important and the adoption of modern way of life is more relevant. With regard to the Geography of the disease, the male pattern is usually strongly spatially structured due to the location of industrial plants, while for females the spatial pattern tends to reflect the degree of urbanization in the region (e.g. Biggeri *et al.*, 2000).

In this paper we review the models proposed to analyze bivariate disease pattern and address the inference on shared and specific risk components. We applied the models to the analysis of mortality for lung cancer by gender in Tuscany among birth cohorts 1905-1940.

(*) This research was partially supported by PRIN 2002134337 and PRIN 2004137478. We are grateful to Dr. Mariangela Vigotti (University of Pisa) and Dr. Elisabetta Chellini (CSPO, Florence and Regional Mortality Register) for having kindly made available the data used in the present work.

2. Data

Lung cancer death certificates from 1971 to 1999 of males and females resident in the 287 municipalities of the Tuscany Region (Italy) were considered. Data were made available by the Tuscany Regional Government (Vigotti *et al.*, 2001) and by the Regional Mortality Register for the period 1995-99. Death counts and corresponding population denominators for each municipality were cross classified by 18 age classes (0-4, ..., 85 and more) and 6 calendar periods (1971-74, ..., 1995-99). On the Lexis diagram we choose six birth cohorts (1905-15, ..., 1930-40) corresponding to people aged between 35 and 64 years at the beginning of the study period. The expected number of cases in each municipality have been obtained applying the age-specific reference rates calculated by an age-cohort model fitted to the aggregate regional data.

3. Space-time bivariate models

We describe the space-time pattern of mortality risk for lung cancer males and females for the whole region from the cohort born on 1905-15 to the cohort 1930-40 using hierarchical Bayesian models with structured random effects on space and time dimensions (Knorr-Held 2000, Lagazio *et al.* 2003). The number of observed cases in the i -th area ($i = 1, \dots, 287$), j -th cohort ($j = 1905-15, \dots, 1930-40$) and k -th sex ($k = M$ male, F female) O_{ijk} are assumed to follow a Poisson distribution with mean $E_{ijk}\theta_{ijk}$, where E_{ijk} indicates the expected number of cases under internal indirect standardization and θ_{ijk} the relative risk. A random effects model is assumed for the logarithm of the relative risk

$$\log(\theta_{ijk}) = \alpha_k + u_{ik} + v_{ik} + p_{jk} \quad (1)$$

where α_k represents a sex-specific intercept and is assumed to follow an improper uniform prior, u_{ik} a spatially structured term by area and sex, v_{ik} a spatially unstructured term by area and sex, p_{jk} a time structured term by cohort and sex.

We explored several alternative specifications for u_{ik} , v_{ik} and p_{jk} , considering various combinations of shared components.

First we consider models with terms which are either common or not: no common terms (1. Us Vs Ps), common cohort effects (2. Us Vs P), common heterogeneity (3. Us V Ps) or common clustering (4. U Vs Ps), common clustering and heterogeneity (5. U V Ps), common clustering and cohort (6. U Vs P), common heterogeneity and cohort (7. Us V P), common heterogeneity clustering and cohort (8. U V P).

We then define 5 more models in which the two spatial random terms (heterogeneity and clustering) are decomposed into a shared and an unshared effect (Knorr-Held and Best, 2001). The cohort effect is introduced in these models only as gender-specific due to the bad fit of models with common cohort term. Following the previous notation the models are (9. Us V Vs Ps), (10. U Us Vs Ps), (11. U V Vs Ps), (12. U Us V Ps) and (13. U Us V Vs Ps).

Taking the last model as reference, since it is the most complex and best fitting one, we can represent the clustering terms for males as $u_{iM} = u_i \times \delta + u_{im}$ and for females as $u_{iF} = u_i/\delta + u_{if}$; the heterogeneity terms for males as $v_{iM} = v_i \times \omega + v_{im}$ and for females as $v_{iF} = v_i/\omega + v_{if}$ where v_i and u_i represent the shared heterogeneity and clustering components and v_{im} v_{if} u_{im} and u_{if} the unshared ones; δ and ω are Lognormal

distributed and allow the shared component to vary by sex by a constant factor. Since the prior for δ and ω are symmetric around zero on a log-scale, any value for these parameters are as “equally likely” as the reciprocal values a priori. The term v_i has a prior distribution assumed to be Normal $(0, \lambda_v)$. The components v_{im}, v_{if} are modeled in the same way, with precision respectively λ_{vm} and λ_{vf} . Term u_i , the shared clustering component, is modeled, conditionally on $u_{l \sim i}$ terms ($\sim i$ indicates areas adjacent to i -th ones, $l = 1, \dots, 287$), as Normal $(\bar{u}_i, \lambda_u n_i)$ where $\bar{u}_i = \sum_{l \sim i} \frac{u_l}{n_i}$. Again, $u_{ik}, k = m, f$, the unshared clustering components, are modeled in the same way, with precision λ_{uk} . The term p_{jk} represents the effect of the j -th cohort for each sex, and is assumed $p_{jk} \sim \text{Normal}(\bar{p}_{jk}, \lambda_{pk} n_j)$; $\bar{p}_{jk}$ is the mean of the $(j - 1)$ -th and $(j + 1)$ -th terms and n_j equal 2 (or 1 for the extreme cohorts). The hyperprior distributions of the precision parameters $\lambda_v, \lambda_{vk}, \lambda_u, \lambda_{uk}$ and λ_{pk} are assumed to be Gamma $(0.5, 0.0005)$ (Kelsall and Wakefield 1999).

Selecting a suitable model from a large class of plausible models is a difficult task and requires special attention; as the candidate models are not completely nested, the relative merit of a given model should be evaluated following a combined criterion that takes into account data support as measured by the Deviance Information Criterion (DIC) (Spiegelhalter *et al.*, 2002) and inference robustness as quantified by the calibrated Kullback-Leibler divergence (KL) (McCulloch, 1989). Since the goal of the present study is to give a description of the phenomenon under study, we consider the DIC criterion instead of other such as Expected Predictive Deviance (Gelfand and Ghosh, 1998). The DIC is the sum of two components: the posterior expectation of deviance that summarize the fit of the model and the effective number of parameters that represents the complexity of the model.

4. Results and conclusions

The birth cohort effects (Figure 2a) were very remarkable by gender. The epidemic curve showed very different patterns: among males the peak was experienced by the birth cohort born around 1930, the younger cohort being at lower risk. Among females the relative risk was rising almost linearly on the log scale, the younger birth cohort being at higher risk. Models 2, 6, 7, 8 with common cohort pattern were the worst fitting.

The Model 13, the reference model, has both shared and unshared components for heterogeneity and clustering terms. Model 5 with common clustering and heterogeneity provide a very different inference from model 13 and also a lack of fit: the geography of the disease seemed different by gender.

Considering jointly data support and inference robustness two slightly parsimonious models can be considered, model 10 and model 12 (Figure 1). The clustering term had a shared component (Figure 2b) for both models, highlighting the historical urban structure of the Tuscany Region which is strongly spatially structured, and differ by the heterogeneity term. Model 10 includes an unshared heterogeneity component which had a nice interpretation because it marked the big cities at higher risk among females (Figure 2c) and small cities with industrial plants among males (Figure 2d).

Figure 1: Calibrated Kullback Leibler divergence versus DIC measure (ref model is 13).

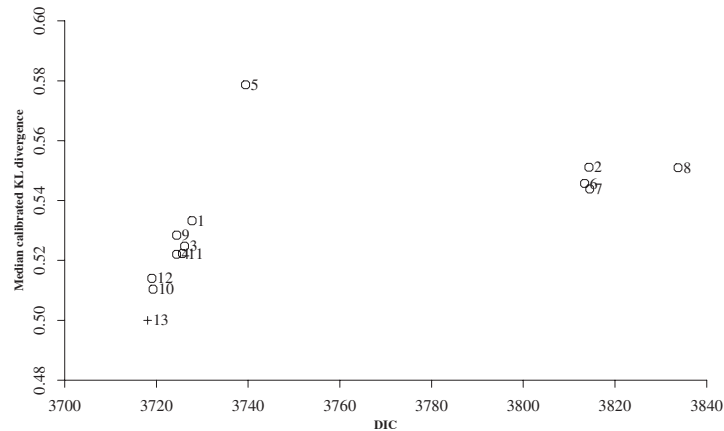
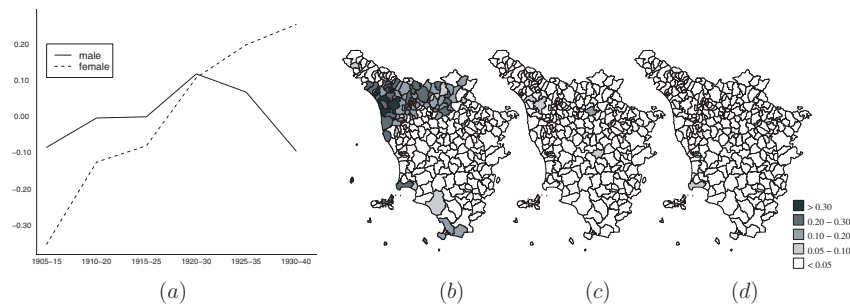


Figure 2: Gender-specific heterogeneity, cohort and shared clustering terms (model 10).



References

- Biggeri A., Marchi M., Lagazio C., Boehning D., Martuzzi M. (2000) Non-parametric Maximum Likelihood Estimators for Disease Mapping, *Statistics in Medicine*, 19, 2539–2554.
- Gelfand A.E., Ghosh S.K. (1998) Model choice: a minimum posterior predictive loss approach, *Biometrika*, 85, 1–11.
- Kelsall J. E., Wakefield J. C. (1999) Discussion of “Bayesian Models for Spatially Correlated Disease and Exposure Data”, in: *Bayesian Statistics 6*, Bernardo *et al.* (eds.), 151, New York Oxford University Press.
- Knorr-Held L. (2000) Bayesian modelling of inseparable space-time variation in disease risk, *Statistics in Medicine*, 17-18, 2555–2568.
- Knorr-Held L., Best N. G. (2001) Bayesian modelling of inseparable space-time variation in disease risk, *JRSS A*, 164, 73–85.
- Lagazio C., Biggeri A., Dreassi E. (2003) Age-Period-Cohort Models and Disease Mapping, *Environmetrics*, 14, 475–490.
- McCulloch R. E. (1989) Local Model Influence, *JASA*, 84, 473–478.
- Spiegelhalter D. J., Best N. G., Carlin B. P., van der Linde A. (2002) Bayesian measures of model complexity and fit (with discussion), *JRRS B*, 64, 583–639.
- Vigotti M. A., Biggeri A., Dreassi E., Protti M. A., Cislighi C. (2001) *Atlas of mortality in Tuscany 1971-94*, Edizioni Plus: Università degli Studi di Pisa.