Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2006/05

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Polytomous disease mapping to detect uncommon risk factors for related diseases

Emanuela Dreassi

Università degli Studi
di Firenze

Polytomous Disease Mapping to detect uncommon risk factors for related diseases

Emanuela Dreassi[*1]

[1] Department of Statistics 'G. Parenti', University of Florence, Viale Morgagni 59, I 50134, Florence, Italy.

Summary

This paper introduces a statistical model for jointly analysing the spatial variation of incidences of three (or more) diseases with common and uncommon risk factors.

We have considered the mortality data (from 1990 to 1994) relative to oral cavity, larynx and lung cancers in 13 age groups of males, in the 287 municipalities of Region of Tuscany (Italy). All these pathologies share smoking as a common risk factor; furthermore, two of them (oral cavity and larynx cancer) share alcohol consumption as a risk factor. All studies suggest that smoking and alcohol consumption are the major known risk factors for oral cavity and larynx cancers; nevertheless, in this paper we investigate the possibility of there being other different risk factors for these diseases or even a different interaction between smoking and alcohol risk factors.

A logit model for multinomial responses (multinomial logit or polytomous logit model) was used to model deaths for the different diseases. For each municipality and age-class we estimated the probabilities of death for each cause (the response probabilities). Lung cancer acts as the baseline category. The log odds are decomposed additively into shared (common to oral cavity and larynx diseases) and specific structured spatial variability terms, unstructured unshared spatial terms and an age-group term to adjust the crude observed data for effects that can be attributed to age. We estimated disease specific spatially structured effects; these are considered as latent variables denoting disease-specific risk factors.

Results show that oral cavity and larynx cancer have different spatial patterns for residual risk factors which are not the typical ones already considered such as smoking habits and alcohol consumption. But, probably, these patterns are due to different spatial interactions between smoking habits and alcohol consumption for the first and the second disease.

Key words: Joint disease mapping, proportional mortality model, shared component model, Hierarchical Bayesian model, Polytomous logit model, oral cavity cancer, larynx cancer, lung cancer.

1 Introduction

A great amount of the literature concerns disease mapping, as the statistical analysis of geographical patterns of disease. Any spatial variations may be explained by different risk factors, therefore disease mapping allows us to formulate hypotheses regards their aetiology. Over recent years there has been increased interest in joint disease mapping: joint statistical modelling of several diseases on the same spatial location, with different and common aetiologies. Joint analysis permits us to highlight common and uncommon geographical patterns of risk and obtain more precise and convincing results.

The first attempts reported, Langford *et al.* (1999) and Leyland *et al.* (2000), introduced joint spatial analysis for two diseases using a multilevel approach. An ecological regression approach has also been suggested by Bernardinelli *et al.* (1997), where disease rates from a second disease enter as a covariate. However, the joint modelling approach seems to be more naive, because diseases enter as response variables in relation to unobserved latent risk factors. Joint modelling following a Multivariate Gaussian Markov random field has been proposed (see Gelfand and Vounatsou 2003 and Jin *et al.*, 2005) as well as a *shared component model* (Knorr-Held and Best, 2001). This latter class of models and its extension

* Corresponding author: e-mail: dreassi@ds.unifi.it, Phone: +39 055 4237219, Fax: +39 055 4223560

to more than two diseases (Held *et al.*, 2005) allow the linear predictor to be decomposed into shared and disease-specific spatial variability components.

Dabney and Wakefield (2005) proposed a *proportional mortality model* (see Breslow and Day, 1987) to the joint mapping of two diseases when the population at risk is unknown. Instead of adopting a Poisson model with expected cases as offset, their model uses a simultaneous estimation of age and spatial effects that should be preferred to the Poisson one, since it includes variability in the age estimates. The model assumes proportionality, as the model sum over strata population and considers only a single parameter per confounder (i.e. age, sex, race) and one set of parameters for each area, without considering their interaction. Inference can be made on the differences in log relative risks without knowledge of the population counts of those at risk.

The proportional mortality model and the shared component model both highlight similarity and dissimilarity on spatial patterns. Considering the extension of the proportional mortality model to more than two diseases, we defined a polytomous logit model (for example, see Agresti, 2002), where a disease is considered as reference category, and for each predictor we adopted the shared component model formulae.

We considered three diseases: oral cavity, larynx and lung cancer. The incidence of lung cancer as a surrogate for the smoking risk factor represents the reference category for the polytomous logit model. As smoking and alcohol consumption are the two most important established risk factors for oral cavity and larynx cancers (i.e. Elwood *et al.*, 1984), we investigated the possibility of there being other different risk factors or perhaps a different interaction between smoking and alcohol consumption risk factors for oral cavity and larynx cancers (see Held *et al.*, 2005 for the German example on oral cavity, oesophagus, larynx and lung cancers).

The paper is organized as follows. Section 2 describes the data and presents the separate analysis using binomial logistic models. Section 3 introduces the joint analysis with polytomous logistic model. Results obtained are showed in Section 4. In Section 5 we discuss the proposed model and results.

2 Data and single analysis

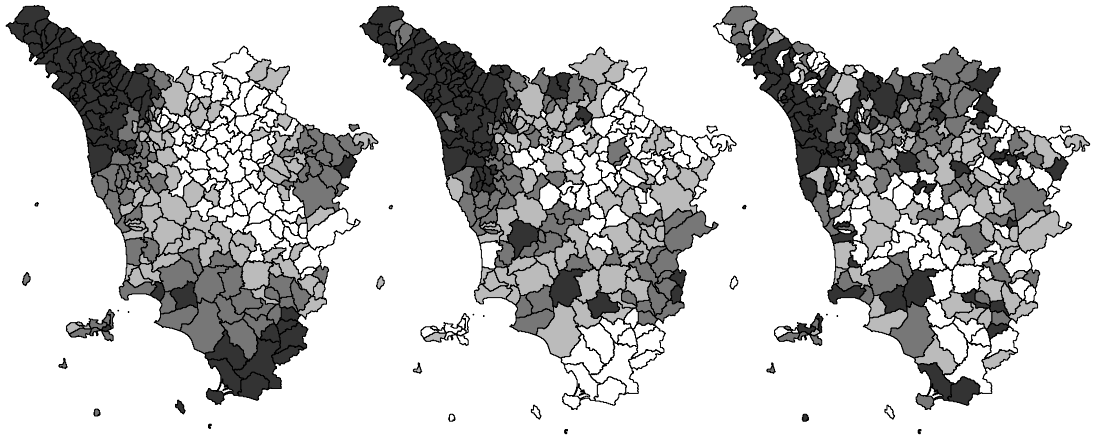
We took into account the death certificates drawn up between 1990 and 1994 that indicated oral (lip, oral cavity and pharynx), larynx and lung (trachea, bronchus and lung) cancers as the causes of death in males resident in the 287 municipalities of the Tuscany Region (Italy). These data were made available to us by the Tuscany Regional Government (Vigotti *et al.*, 2001). Death counts for each cause $k = 1, 2, 3$ and municipality $i = 1, \ldots, 287$ were cross-classified by 18 age-classes, due to sparseness of data the $j = 1, \ldots, 13$ age-classes considered are 25-29, ..., 80-84, 85 plus.

We first consider a single disease analysis for counts death using a binomial model and considering the population at risk. Let y_{ijk} denote the number of death cases for k-th disease ($k = 1, \ldots, K$), in age-group j-th ($j = 1, \ldots, J$) and area i-th ($i = 1, \ldots, I$). We assume that each y_{ijk} follows a binomial distribution with parameters n_{ij} and binomial probability π_{ijk}, where n_{ij} represent persons-years at risk in area i-th and age-group j-th for all the diseases. The likelihood for the entire data is the corresponding product of binomial terms. We follows standard model considering a logit link for π_{ijk}

$$\text{logit}(\pi_{ijk}) = \eta_{ijk} = \alpha_k + a_{jk} + u_{ik} + v_{ik} \tag{1}$$

where each log-odds is decomposed additively into constant, age-group and spatially structured and unstructured effects. α_k represents a cause-specific intercept, as an overall risk level and on the Besag *et al.* (1991) philosophy, using Markov random field models in order to cope the spatial structure, u_{ik} a spatially structured term by area and cause. Terms v_{ik} represent a spatially unstructured term by area and cause and a_{jk} a time-structured term by age and cause.

For α_k we assume a flat non-informative prior distribution. The v_{ik} heterogeneity terms have a prior distribution assumed to be Normal $(0, \lambda_{vk})$ each $k = 1, 2, 3$. Terms u_{ik}, the clustering components, are modeled $k = 1, 2, 3$ conditionally on $u_{l \sim ik}$ terms ($\sim i$ indicates areas adjacent to i-th ones, $l = 1, \ldots, 287$



oral cavity *larynx* *lung*

Fig. 1 Oral, larynx and lung cancer odds probabilities (quantiles), age class 75–79, Tuscany, 1990–94.

and n_i their number), as Normal $(u_{\bar{i}k}, \lambda_{uk} n_i)$ where $u_{\bar{i}k} = \sum_{l \sim i} \frac{u_{lk}}{n_i}$. The term a_{jk} represents the effect of the j-th age for each cause, and is assumed $a_{jk} \sim$ Normal $(a_{\bar{j}k}, \lambda_{ak} n_j)$; $a_{\bar{j}k}$ is the mean of the $(j-1)$-th and $(j+1)$-th terms and n_j equal 2, for the extreme age classes n_j equal 1 and $a_{\bar{j}k}$ is the $(j+1)$-th or $(j-1)$-th term. The hyperprior distributions of the precision parameters λ_{vk}, λ_{uk} and λ_{ak} are assumed to be Gamma $(0.5, 0.0005)$ as suggested by Kelsall and Wakefield (1999).

Figure 1 describes the quantiles of odds probability from binomial model single analysis for oral cavity $\exp(\eta_{i\,11\,1})$, larynx $\exp(\eta_{i\,11\,2})$ and lung $\exp(\eta_{i\,11\,3})$ cancer for age-class 75-79. These maps show that all the diseases share a common spatial pattern, reflecting smoking habits, with highest rates in the northwestern areas.

Figure 2 shows the multiplicative age effects on odds probabilities for the three diseases $\exp(a_{j1})$, $\exp(a_{j2})$ and $\exp(a_{j3})$; we note the different epidemic curve for the diseases according to age.

3 Joint analysis: the Polytomous logit model

Let y_{ijk} denote the number of death cases for k-th disease ($k = 1, \ldots, K$), in age-group j-th ($j = 1, \ldots, J$) and area i-th ($i = 1, \ldots, I$). We assume that $y_{ij} = (y_{ij1}, \ldots, y_{ijk}, \ldots, y_{ijK})\prime$ follows a multinomial distribution (as in Knorr-Held *et al.*, 2002 when cumulative logit model has been used to analyzing stage-specific for a single disease) with parameters m_{ij} and probability vector $\pi_{ij} = (\pi_{ij1}, \ldots, \pi_{ijk}, \ldots, \pi_{ijK})\prime$, where $m_{ij} = \sum_{k=1}^{K} y_{ijk}$ and $\sum_{k=1}^{K} \pi_{ijk} = 1$.

We consider a polytomous logit model considering the proportional mortality model (see also Dabney and Wakefield, 2005) so we don't need population data; each category probability is modeled as

$$\pi_{ijk} = \phi_{ijk} / \sum_{k=1}^{K} \phi_{ijk} \tag{2}$$

where each log odds, $\log(\phi_{ijk})$, is decomposed additively into constant, age-group and spatial disease-specific effects following a proportionality assumption

$$\log(\phi_{ijk}) = \alpha_k + a_{jk} + u_{ik} + v_{ik} \tag{3}$$

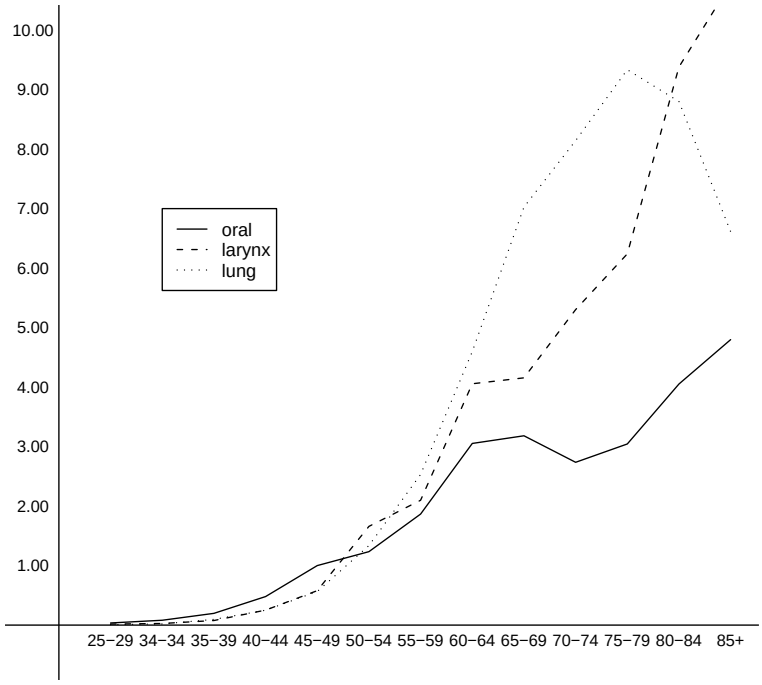


Fig. 2 The different epidemic curve by age for oral cavity, larynx and lung cancers.

where α_k represents a disease-specific intercept, u_{ik} a spatially structured term by area and disease, v_{ik} a spatially unstructured term by area and disease, a_{jk} a time-structured term by age and disease. We consider a model with the structured spatial terms (clustering) has been decomposed into a shared and a disease-specific effect (Knorr-Held and Best, 2001).

In our example we consider three disease, $K = 3$: oral cavity, larynx and lung cancers. Lung cancer is the reference category, so that we set α_3, a_{j3} (for each age-class $j = 1, \ldots, 13$), u_{i3} and v_{i3} (for each municipalities $i = 1, \ldots, 287$) equal to zero as constraint for identifiability. Inference has been made on the differences in log relative risks between oral cavity and lung cancer and larynx and lung cancer, without knowledge of the population counts.

We can represent each clustering terms for oral cavity and larynx cancer respectively as

$$u_{i1} = u_i \times \delta + us_{i1} \qquad \text{and} \qquad u_{i2} = u_i/\delta + us_{i2} \tag{4}$$

where u_i represent the shared clustering components and us_{i1} and us_{i2} the specific ones; δ is Lognormal distributed and they allow the shared component to vary by cause by a constant factor. Since the prior for δ is symmetric around zero on a log-scale, any value for these parameters are as "equally likely" as the reciprocal values a priori. For α_k we assume a flat non-informative prior distribution. The v_{i1} and v_{i2} represent unstructured spatial terms: v_{i1} and v_{i2} have a prior distribution respectively assumed to be Normal $(0, \lambda_{v1})$ and Normal $(0, \lambda_{v2})$. 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, $l = 1, \ldots, 287$ and n_i their number), as Normal $(\bar{u_i}, \lambda_u n_i)$ where $\bar{u_i} = \sum_{l\sim i} \frac{u_l}{n_i}$. Again, us_{ik}, $k = 1, 2$, the specific clustering components, are modeled in the same way, with precision λ_{uk}. The term a_{jk} represents the effect of the j-th age for each cause, and is assumed $a_{jk} \sim$ Normal $(\bar{a_{jk}}, \lambda_{ak} n_j)$ (a first order random walk prior); $\bar{a_{jk}}$ is the mean of the $(j-1)$-th and $(j+1)$-th terms and n_j equal 2, for the extreme age classes n_j equal 1 and $\bar{a_{jk}}$ is the $(j+1)$-th or

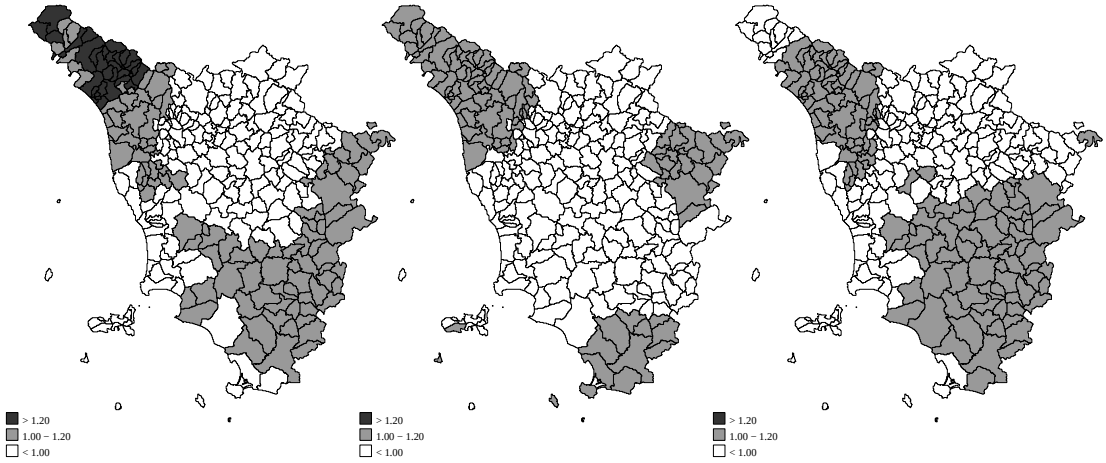


Fig. 3 Oral and larynx shared and specific structured exponential spatial terms.

$(j-1)$-th term. The hyperprior distributions of the precision parameters λ_{vk}, λ_u, λ_{uk} and λ_{ak} are assumed to be again Gamma $(0.5, 0.0005)$ as suggested by Kelsall and Wakefield (1999).

For models described in Section 2 and 3, the marginal posterior distributions of the parameters of interest were approximated by Monte Carlo Markov Chain methods. We have made use of WinBUGS software (Spiegelhalter *et al.*, 2000) in order to perform the MCMC analysis. For each model we have run two independent chains; checks for achieved convergence of the algorithm were performed following Gelman and Rubin (1992).

The age effects from these models do not have an interesting and direct interpretation; their inclusion in the model is only to decontaminate an age-confounding effect. Interest is focused on the estimate of disease-specific spatially structured effects because these are considered as latent variables denoting disease-specific risk factors.

4 Results

Figure 3 shows shared $\exp(u_i)$ and disease specific $\exp(us_{i1})$ and $\exp(us_{i2})$ structured spatial effects. The shared structured spatial effects represent alcohol consumption. This component is much higher in the northwestern and southeastern parts of Tuscany. As the category reference is lung cancer, areas with lower values could be due to occupational risk factors for this disease, and areas with higher values where no smoking habits or occupational risks are present, but only alcohol consumption. The mean of the posterior distribution for the scaling parameter δ is 1; this indicates that unobserved risk factors (alcohol) that are common for oral cavity and larynx cancer are associated on the same magnitude for these tumors.

The disease-specific structured spatial effects represent differences between oral cavity and larynx cancer. Disease-specific spatial terms with values higher than 1 suggest the possibility of having additional, but uncommon, risk factors for each disease. We noted that when considering smoking and alcohol consumption risk factors, oral cavity and larynx have different specific component spatial patterns due to residual risk factors or, more convincingly, a different interaction between smoking habits and alcohol consumption.

5 Conclusions and discussion

In this paper we have introduced a joint spatial analysis of three diseases with common and uncommon risk factors to detect the presence of further risk factors.

We extended the proportional mortality models to more than two diseases and we adopted a shared component model to define their liner predictors.

The model proposed by Held *et al.* (2005) for joint disease mapping is perhaps more 'natural' and 'elastic' than our model; however, the latter gives clear advantages since it permits us to analyse data on mortality without having knowledge of the population at risk and to consider variability on age effect estimates in the model.

The identification of a specific spatial pattern risk for a disease might suggest not only the presence of uncommon risk factors but, probably, a different interaction between the known risk factors: smoking habits and alcohol consumption. In fact, an interaction between smoking and alcohol consumption, suggesting a synergistic effect, has been found repeatedly for oral cavity and larynx cancer and perhaps we can now demonstrate that this is probably slightly different.

Nevertheless, the study relies on aggregate data and no definitive causal conclusions can be drawn, since ecological biases cannot be excluded. Moreover, it may be important to consider time dimension for a larger dataset or to consider the difference on latency-time between the first exposure to risk factors and death due to these diseases.

Acknowledgements The research was partially supported by PRIN 2002134337 and PRIN 2004137478. The author expresses thanks to Dott.ssa Mariangela Vigotti from University of Pisa for having kindly made available the data used in the present work.

References

Agresti A. (2002) *Caegoriacal Data Analysis*, 2^{nd} edition. Wiley series in probability and statistics, John Wiley & Sons, Inc.

Bernardinelli, L., Pascutto, C., Best, N.G. and Gilks, W.R. (1997) Disease mapping with errors in covariates. *Statistics in Medicine* **16**, 741–752.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43**, 1–59.

Breslow, N. and Day, N.E. (1987) Statistical Methods in cancer research, volume 2. *The analysis of cohorts studies*. Scientific publications n. 82, Lyon: International Agency for Research on Cancer.

Dabney, A.R. and Wakefield, J.C. (2005). Issues in the mapping of two diseases. *Statistical Methods in Medical Research* **14**, 83–112.

Elwood, J.M., Pearson, J.C., Skippen, D.H. and Jackson, S.M. (1984). Alcohol, smoking, social and occupational factors in the aetiology of cancer of the oral cavity, pharynx and larynx. *International Journal of Cancer* **34**, 603–612.

Gelfand, A. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* **4**, 11–25.

Gelman, A. and Rubin, D.R. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science* **7**, 457–511.

Held, L., Natário, I., Fenton, S.E., Rue, H. and Becker, N. (2005). Towards joint disease mapping. *Statistical Methods in Medical Research* **14**, 61–82.

Jin, X., Carlin, B.P. and Banerjee, S. (2005). Generalized Hierarchical Multivariate CAR Models for Areal Data. *Biometrics* **61**, 4, 950–961.

Kelsall, J.E. and Wakefield, J.C. (1999). Discussion of "Bayesian Models for Spatially Correlated Disease and Exposure Data", by Best *et al.*, in *Bayesian Statistics 6*, Bernardo *et al.* (eds.), Ney York Oxford University Press, 151.

Knorr-Held, L. and Best, N. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society Series A (Statistics in Society)* **164**, 73–86.

Knorr-Held, L., Raßer, G. and Becker, N. (2002). Disease Mapping of stage-specific cancer incidence data. *Biometrics* **58**, 492–501.

Langford, I.H., Leyland, A.H., Rasbash, J., Goldstein, H. (1999) Multilevel modelling of the geographical distributions of diseases. *Journal of the Royal Statistics C - Applied Statistics* **48**, 253-268.

Leyland, A.H., Langford, I.H., Rasbash, J. and Goldstein, H. (2000). Multivariate spatial models for event data. *Statistics in Medicine* **19**, 17-18, 2469–2478.

Spiegelhalter, D.J., Thomas, A., Best, N. and Lunn, D. (2002). *WinBUGS User Manual, Version 1.4*. (Online user manual, http://www.mrc-bsu.cam.ac.uk/bugs: accessed 20 January 2004).

Vigotti, M.A., Biggeri, A., Dreassi, E., Protti, M.A., Cislaghi, C. (2001). *Atlas of mortality in Tuscany 1971-94*. Edizioni Plus: Università degli Studi di Pisa.