Geospatial Health 3(1), 2008, pp. 81-89

A statistical approach to rank multiple priorities in Environmental Epidemiology: an example from high-risk areas in Sardinia, Italy

Dolores Catelan[1,2], Annibale Biggeri[1,2]

[1]*Department of Statistics "G. Parenti", University of Florence, Florence, Italy;* [2]*Biostatistics Unit, ISPO Cancer Prevention and Research Institute, Florence, Italy*

Abstract. In Environmental Epidemiology, long lists of relative risk estimates from exposed populations are compared to a reference to scrutinize the dataset for extremes. Here, inference on disease profiles for given areas, or for fixed disease population signatures, are of interest and summaries can be obtained averaging over areas or diseases. We have developed a multivariate hierarchical Bayesian approach to estimate posterior rank distributions and we show how to produce league tables of ranks with credibility intervals useful to address the above mentioned inferential problems. Applying the procedure to a real dataset from the report "Environment and Health in Sardinia (Italy)" we selected 18 areas characterized by high environmental pressure for industrial, mining or military activities investigated for 29 causes of deaths among male residents. Ranking diseases highlighted the increased burdens of neoplastic (cancerous), and non-neoplastic respiratory diseases in the heavily polluted area of Portoscuso. The averaged ranks by disease over areas showed lung cancer among the three highest positions.

Keywords: Environmental Epidemiology, hierarchical Bayesian model, rank, multiple comparisons.

Introduction

Descriptive geographical studies of areas at high-risk for environmental pressure aim to screen the health status of populations, to identify priorities for public health interventions or to suggest further analytical studies (Elliott et al., 2000). This kind of study is usually carried out on a predefined number of areas at the national (see for example Mitis et al., 2005) or the regional level (see for example Biggeri et al., 2006). For each area a large number of diseases are evaluated, eventually by gender, in terms of mortality or morbidity. The results are presented as long lists of relative risks and the problem of multiple comparisons in interpreting p-values is usually not addressed (Rothman, 2002). Alternatively, several procedures, put forward by Catelan et al. (2006), can be considered such as:

(i) build a league-table based on ranks and associated intervals (Goldstein and Spiegelhalter, 1996);
(ii) build a Q-Q plot of test statistics with guide rails (Carpenter et al., 1997; Law et al., 2001); or
(iii) control the positive-false discovery rate (Storey, 2003).

The first approach is coherent with Bayesian statistical modelling of disease risk and allows summary description of the evidence and related uncertainty. The second method is attractive since it allows non-parametric exploration of few extreme risk values by unit or by disease code. The last tactic is more decision-oriented, and can be used to evaluate research hypotheses or public health scenarios.

Ordered statistics have been introduced to the purpose of institutional performance assessment in

Corresponding author:
Dolores Catelan
Department of Statistics "G. Parenti"
University of Florence
Viale Morgagni, 59 - 50134 Florence, Italy
Tel. +39 055 4237472; Fax +39 055 4223560
Email: catelan@ds.unifi.it

education and in health service research (Goldstein and Spiegelhalter, 1996) but so far no extension has been proposed in environmental epidemiology, except for some mortality atlases (see for example Vigotti et al., 2001). In this work, motivated by the "Environment and Health in Sardinia (Italy)" report (Biggeri et al., 2006), we have focused on the "league-table" approach. The statistical method used for estimating ranks is described in the methods section. Results based on real examples and a discussion close the paper.

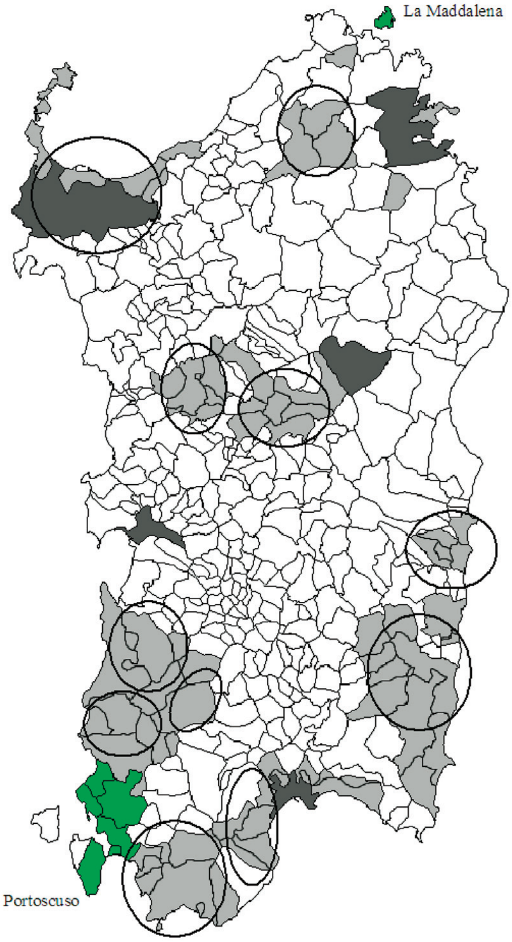

Fig. 1. Map of Sardinia (Italy) with the 18 investigated areas. The circles show the industrial, mining and military areas. The rural districts are shown in light grey and the metropolitan areas (Cagliari, Sassari, Oristano, Nuoro, Olbia), in dark grey. The industrial area of Portoscuso and the military area of La Maddalena are marked in green.

Materials and methods

Motivating example

The report "Environment and Health in Sardinia (Italy)" (Biggeri et al., 2006) was committed by the local government of the region to screen the health status of resident populations in *a priori* identified areas being at high environmental pressure for the presence of industrial, mining and military activities (18 areas with 917,977 individuals in total according to the 2001 census, i.e. 56% of Sardinia's population; Fig. 1). For each area, separately for males and females, mortality data for the period 1981-2001 and hospital discharge data for the period 2001-2003 were analyzed. A series of statistical estimators was applied such as crude and age-standardized rates, standardized mortality ratios (SMR), SMR adjusted also for material deprivation, Bayesian SMR and posterior probabilities of risk greater than the regional average, model-based calendar period and birth cohort relative risks. Overall we analysed 84 disease codes, 36 for mortality and 48 for hospital admissions, stratified by 18 areas, resulting in 1,512 relative risk estimates for each sex. We are thus in a "large table" context (Carpenter et al., 1997) in the following: we could consider, without loss of generality, only mortality data among males over the period 1997-2001 and 29 non-overlapping causes of death (the list of causes according to the 9th version of International Classification of Disease (ICD-9) is reported in Tables 1a and 1b). Using internal indirect standardization, the data were collapsed to area-observed and expected death counts. All expected counts were adjusted for material deprivation (Grisotto et al., 2007). The statistical task was to summarize 522 comparisons (18 areas and 29 diseases) with the reference population.

Statistical analysis

Generally speaking a vector of ranks $\boldsymbol{R}=(R_1,\ldots,R_k,\ldots,R_K)$ can be defined as the set of ranks

$$R_k = \text{rank}(\vartheta_k) = \Sigma_{k'=1}^{k} I[\vartheta_k \geq \vartheta_{k'}]$$

where $I(\cdot)$ is the indicator function and ϑ_k the unit specific parameter of interest.

In our analysis we were interested in comparing relative risks of J diseases ($j = 1,\ldots,J$) for I areas ($i = 1,\ldots I$). Therefore we have produced a matrix $\boldsymbol{P_{IxJ}}$ of ranks whose rows represent disease profiles by area, i.e. area vectors $\boldsymbol{P_{i.}}=(P_{i1},\ldots,P_{ij},\ldots,P_{iJ})$ with

$$P_{ij} = \text{rank}(\vartheta_{ij}) = \Sigma_{j'=1}^{I} I[\vartheta_{ij} \geq \vartheta_{ij'}]$$

and a matrix $\boldsymbol{S_{IxJ}}$ of ranks whose columns represent area signatures by disease, i.e. vectors $S_{.j}=(S_{1j},\ldots,S_{ij},\ldots,S_{Ij})$ with

$$S_{ij} = \text{rank}(\vartheta_{ij}) = \Sigma_{i'=1}^{I} I[\vartheta_{ij} \geq \vartheta_{i'j}]$$

Marginalizing over columns in matrix $\boldsymbol{P_{IxJ}}$ of disease profiles we can have an overall disease ranking, while marginalizing over rows in matrix $\boldsymbol{S_{IxJ}}$ of area signatures we can obtain an overall area ranking.

Let Y_{ij}, the observed number of cases in the i-th

Table 1a. Environment and health in high-risk areas of Sardinia, Italy. List of chosen causes of death with ICD-9 code. Non-neoplastic diseases.

Type of disease	ICD-9 code
Infectious disease	001-139
Diabetes	250
Circulatory disease (excluding coronary disease)	390-459 (excluding 410-414)
Coronary disease	410-414
Respiratory disease (excluding chronic and pneumoconiosis)	460-519 (excluding 490-496)
Respiratory chronic disease	416, 490-496
Pneumoconiosis	500-505
Digestive disease (excluding cirrhosis)	520-579 (excluding 571)
Cirrhosis	571
Disease of the urinary system	580-599
Ill-defined conditions	780-799
Accidents and poisoning	800-999

Table 1b. Environment and health in high-risk areas of Sardinia, Italy. List of chosen causes of death with ICD-9 code. Neoplastic diseases.

Target organ	ICD-9 code
Stomach	151
Colon-rectum	153, 154
Liver	155, 156
Larynx	161
Lung	162
Pleura	163
Bone and soft tissue	170, 171
Melanoma	172
Prostate	185
Testis	186
Urinary bladder	188
Central nervous system	191, 192, 225
Thyroid	193
Non-Hodgkin's lymphoma	200, 202
Hodgkin's lymphoma	201
Multiple myeloma	203
Leukemia	204-208

area and j-th disease, follow a Poisson distribution with mean $E_{ij}\theta_{ij}$, where E_{ij} is the expected number of cases under indirect standardization. The parameter θ_{ij}, is the relative risk for the i-th area and j-th disease. We assumed that for each area $\log(\theta_i)$ be independently drawn from a multivariate normal distribution with vector mean $\mu_{\boldsymbol{i}} = (\mu_{i1},\ldots,\mu_{ij},\ldots,\mu_{iJ})$ and JxJ covariance matrix Σ (Assunção and Castro, 2004).

At the third level of the hierarchy we chose a non-informative normal prior distribution with zero mean and precision 0.0001 for each μ_{ij} entry and a Wishart(Ω,ν) distribution for the covariance matrix Σ, where Ω is a JxJ positive definite parameter matrix and ν, the shape parameter, is $\nu \geq J$ (Carlin and Louis, 2009).

Rank posterior distributions of relative risks are obtained from Markov chain Monte Carlo (MCMC) runs (Gilks et al., 1996). The rank of the parameter of interest is computed at each iteration. The MCMC runs approximate the joint cumulative posterior distribution of relative risks F(θ|Y) and hence it is possible to approximate the posterior distribution $[R_k|Y]$ ∀k and its summaries (for example the posterior mean $E[R_k|Y]$ as point estimate of the rank). The posterior mean of the rank is usually not an integer and it is shrunken toward the mid-rank.

Posterior ranks and posterior distributions are obtained:

(i) by disease for a given area, i.e. for disease profiles $\boldsymbol{P_{i.}}=(P_{i1},\ldots,P_{ij},\ldots,P_{iJ})$
$\hat{P}_{ij} = E[P_{ij}|Y] = \Sigma_{j'=1}^{J} F(\vartheta_{ij} \geq \vartheta_{ij'}|Y)$;

(ii) by area for a given disease, i.e. for area signatures $\boldsymbol{S_{i.}}=(S_{i1},\ldots,S_{ij},\ldots,S_{iJ})$
$\hat{S}_{ij} = E[S_{ij}|Y] = \Sigma_{i'=1}^{I} F(\vartheta_{ij} \geq \vartheta_{i'j}|Y)$;
(not considered in the example because the 18 areas are too different in terms of environmental risk factors and the classification of areas by single disease is meaningless);

(iii) by disease marginalizing over areas, i.e. for $\boldsymbol{P}_{..}=(P_{.1},\ldots,P_{.j},\ldots,P_{.J})$;

(iv) by area marginalizing over diseases, i.e. for $\boldsymbol{S}_{..}=(S_{1.},\ldots,S_{i.},\ldots,S_{I.})$.

All computations were performed with WinBugs14 (Spiegelhalter et al., 2003).

Results

Ranking by disease for a given area

Here we report the results for only two areas: Portoscuso, an industrial area in the south of the region in which smelters and foundries (aluminum, lead and zinc), power plants and mines (coal, lead and zinc) are present and La Maddalena, a military area with naval army shipyards and a nuclear submarine base.

Figure 2 shows the posterior ranks of 29 diseases for Portoscuso area. The first three ranks are relative to pneumoconiosis, respiratory diseases and lung cancer and the last three are relative to diabetes, prostate cancer and ill-defined conditions[1]. For these diseases credibility intervals (CrIs) do not overlap indicating a very clear disease profile with higher risk for neoplastic (cancerous) and non-neoplastic affections of the respiratory system and lower risk for the few abovementioned causes. The first position for pneumoconiosis is not surprising because of the occupational exposure of miners.

The classification of the 29 selected causes of deaths for the military area of La Maddalena is reported in Figure 3. Rank 1 is relative to pleural tumours but it is very imprecise with CrI which widely overlaps with the CrIs of other diseases. Only the intervals for non-Hodgkin lymphoma (rank 2) and circulatory diseases (rank 3) are distinguishable from the last one, respiratory diseases, at rank 29.

Ranking by disease marginalizing over area

When we classified the 29 diseases, marginalizing over the 18 area profiles (Fig. 4), diabetes ranks first

[1] more formally the International Classification of Diseases 9th revision code 780-799 "Signs, symptoms and ill-defined conditions" which contains in the basic tabulation list conditions such as pyrexia of unknown origin ICD-9 780.6, symptoms involving heart 785.0-785.3, renal colic 788.0, retention of urine 788.2, abdominal pain 789.0, senility without mention of psychosis 797, sudden infant death syndrome 798.0, respiratory failure 799.1 (http://www.who.int/classifications/en/).

followed by melanoma and lung cancer. The intervals do not overlap with those of the diseases in the last positions of the classification (thyroid cancer, cirrhosis and non-Hodgkin lymphoma). If the risk for a given disease would have been consistently greater/lower than the regional mean in all the screened areas than that disease, it would have occupied the first/last position in the classification when marginalizing over areas. Otherwise, when only some diseases would have changed in only some areas, their marginal position would widely overlap with those of other diseases. In our example, the overall evidence of altered risk is limited to a few diseases, notably lung cancer.

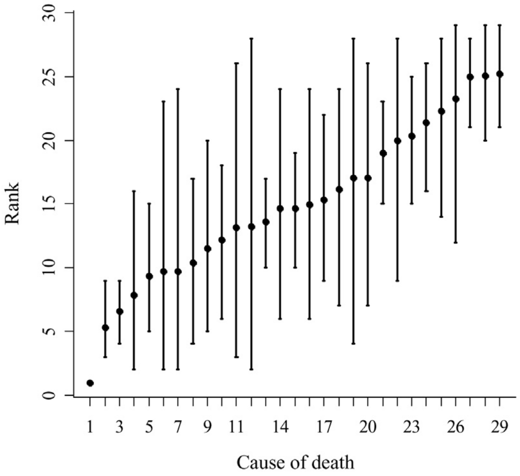

Fig. 2. Mortality in males in the period 1997-2001 in the Portoscuso (Sardinia, Italy) industrial area. Posterior rank estimates and 80% credibility intervals for the selected causes of deaths.

Ranking by areas marginalizing over disease

In Fig. 5, the 18 areas are classified summarizing over the 29 disease signatures. The resulting league-table is flat, i.e. no area outweighs the other. This finding is interesting as, in Sardinia, the selected diseases showed very different signatures. There were no areas with elevated risk for all causes of deaths, i.e. some areas showed altered risks for certain diseases, while other areas did not follow this pattern but instead showed a changed risk for other diseases. This is consistent with the information on the kind of environmental exposures which varies among the different areas.

Table 2. Ranks of diseases reported in Fig. 2. Neoplastic diseases are indicated with N.

Rank	Disease		Rank	Disease	
1	Pneumoconiosis		15	Accidents and poisoning	
2	Respiratory disease (excluding chronic and pneumoconiosis)		16	Non-Hodgkin's lymphoma	N
			17	Liver	N
3	Lung	N	18	Disease of the urinary system	
4	Larynx	N	19	Bone and soft tissue	N
5	Colon-rectum	N	20	Central nervous system	N
6	Hodgkin's lymphoma	N	21	Coronary disease	
7	Pleura	N	22	Multiple myeloma	N
8	Stomach	N	23	Respiratory chronic disease	
9	Urinary bladder	N	24	Cirrhosis	
10	Digestive disease (excluding cirrhosis)		25	Leukemia	N
11	Thyroid	N	26	Melanoma	N
12	Testis	N	27	Prostate	N
13	Circulatory disease (excluding coronary disease)		28	Ill-defined conditions	
14	Infectious disease		29	Diabetes	

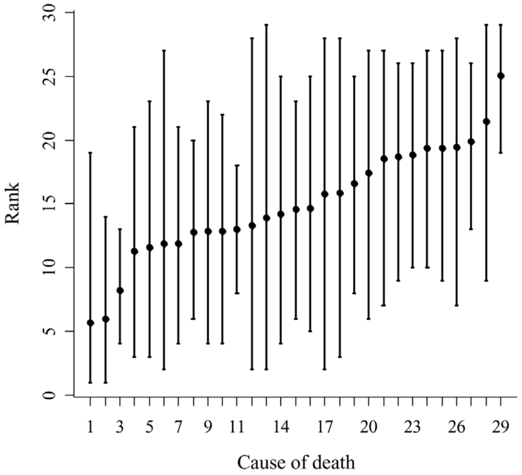

Fig. 3. Mortality in males in the period 1997-2001 in La Maddalena (Sardinia, Italy) military area. Posterior rank estimates and 80% credibility intervals for selected causes of deaths.

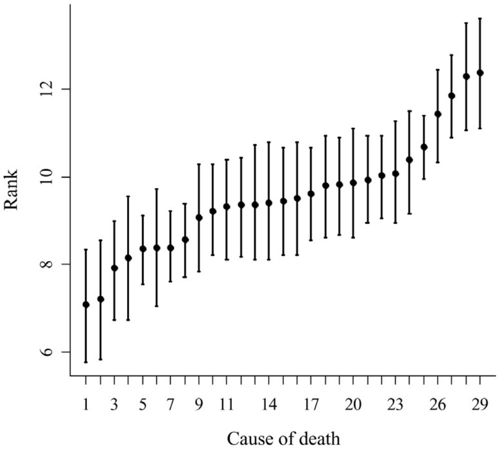

Fig. 4. Mortality in males in the period 1997-2001. Marginal posterior rank estimates and 80% credibility interval for selected causes of deaths, integrated over the 18 areas.

Table 3. Rank of diseases reported in Fig. 3. Neoplastic diseases are indicated with N.

Rank	Disease		Rank	Disease	
1	Pleura	N	16	Larynx	N
2	Non-Hodgkin lymphoma	N	17	Testis	N
3	Coronary disease		18	Hodgkin's lymphoma	N
4	Urinary bladder	N	19	Colon-rectum	N
5	Ill-defined conditions		20	Central nervous system	N
6	Thyroid	N	21	Infectious disease	
7	Cirrhosis		22	Prostate	N
8	Accidents and poisoning		23	Chronic respiratory disease	
9	Leukemia	N	24	Diabetes	N
10	Liver	N	25	Stomach	
11	Circulatory disease (excluding coronary disease)		26	Multiple myeloma	N
12	Bone and soft tissue	N	27	Lung	N
13	Pneumoconiosis		28	Melanoma	N
14	Disease of urinary system		29	Respiratory disease (excluding chronic and pneumoconiosis)	
15	Digestive system (excluding cirrhosis)				

Discussion

For each goal an appropriate table of ranks can be build as summary of the empirical evidence. Focusing on a given area, the diseases classification by higher/lower risk will provide an immediate view to the health profile of the resident population. Local priorities follow. Less frequently, but with notable exceptions, the league-table can be produced for a single disease-code ranking all the screened areas (see, for example, the mortality risk for pleural tumours).

Summarizing health profiles and giving an overall ranking of the areas under surveillance are used to

Table 4. Rank of diseases reported in Fig. 4. Neoplastic diseases are indicated with N.

Rank	Disease		Rank	Disease	
1	Diabetes		15	Chronic respiratory disease	
2	Coronary disease	N	16	Central nervous system	N
3	Lung	N	17	Accidents and poisoning	
4	Multiple myeloma	N	18	Leukemia	N
5	Respiratory disease		19	Ill-defined conditions	
	(excluding chronic and pneumoconiosis)		20	Testis	N
6	Disease of urinary system		21	Bone and soft tissue	N
7	Pleura	N	22	Infectious disease	
8	Urinary bladder	N	23	Coronary disease	
9	Larynx	N	24	Liver	N
10	Stomach	N	25	Colon-rectum	N
11	Prostate	N	26	Pneumoconiosis	
12	Circulatory disease (excluding coronary disease)		27	Thyroid	N
13	Digestive system (excluding cirrhosis)		28	Cirrhosis	
14	Non-Hodgkin's lymphoma	N	29	Hodgkin's lymphoma	N

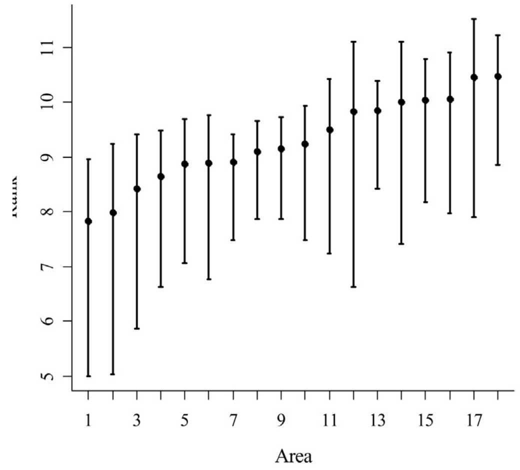

Fig. 5. Mortality in males in the period 1997-2001. Marginal posterior rank estimates and 80% credibility interval for 18 areas integrated over the selected causes of deaths.

identify the most endangered populations and target interventions or environmental remediation. Symmetrically, an overall disease league-table is important in order to check if common exposure patterns were reflected in few specific disease altered risks.

An important point that deserves to be emphasized is that tables of ranks without confidence interval have very limited utility. In fact, the identification of diseases which represent a priority for a given area, or the identification of the set of areas at higher/lower risk for a given disease can be accomplished provided we had applied appropriate filtering of sampling variability. Also for the two marginal tables of ranks we need to consider classification uncertainty. Scrutiny for populations which are generally more compromised may result in a large overlap of the rank confidence intervals: this is informative of disease profiles with few excesses not shared among areas. On the contrary, a clear, highly significant ranking would highlight few areas with risk exceeding for most disease codes. On the other hand, the marginal disease league-table would be blurred if each area would be at high/low risk for a different specific disease, while a clear, statistically significant ranking would be the result of shared risk pattern among areas, with the same excesses/deficits.

Ranks are heavily affected by sampling variability and how to cope with is not straightforward. We specified a multivariate hierarchical Bayesian model and approximate rank posterior distributions from the MCMC runs. We calculated ranks from parameters simulated from the posterior distribution at each iteration of the algorithm. However, the results could be suboptimal if ranks, obtained indirectly by using the posterior means of the parameter of interest, constitute the inferential goal (Shen and Louis, 1998). Laird and Louis (1989) proposed to use posterior expected ranks, while Shen and Louis (1998) introduced a GR estimator which is optimal for estimating the empirical distribution function G and the ranks R. The problem is that using posterior means is poor when posterior variances are heterogeneous.

We left estimated ranks as non-integer values. We did so to communicate the uncertainty on the classification, e.g. the "distance" among ranks (see ordinate axis of Figs. 4 and 5). Some authors (Shen and Louis, 1998) have suggested to rank the ranks.

A multivariate assumption is natural to model a population specific random component for the disease profile. This could be justified in several ways, i.e. genetic composition and susceptibility, and environmental characteristics. Assunção and Castro (2004) used a multivariate model to make inference on the correlation among responses and investigate on the existence of shared risk factors for some diseases. We do not address this point, while stressing area-specific disease profile as the basic object to be ranked.

The use of ranks instead of relative risks could be debatable. However it is an important tool to summarize results and to communicate to lay people and it is not intended to be a substitute for more sophisticated analysis.

In the seminal paper by Goldstein and Spiegelhalter (1996) two kinds of intervals were proposed, namely, conventional interval and overlap interval. The overlap interval averages type I error over all pair-wise comparison and can be extended to multiple comparisons. The conventional confidence interval helps to locate the unit within the overall population distribution. We derived CrIs from the posteriors. Therefore it is more than a simple graphical device and represents information on the whole set of areas at high/low risk. In this sense the Bayesian approach circumvents the issue of multiple comparisons.

Acknowledgements

The present work has been made possible by Ministry of Health ex art. 12 project 2006-387926 Regione Sardegna (2007-2009); Ministry of Education and Scientific Research PRIN 2006131039 "Statistical Methods for Environmental Health Impact Assessment" (2006-2008); European Union, Ministry of Health and AT ESA (QCS-PON ATAS) (2004-2007).

References

Assunção RM, Castro MSM, 2004. Multiple cancer sites incidence rates estimation using multivariate Bayesian model. Int J Epidemiol 33, 508-516.

Biggeri A, Lagazio C, Catelan D, Pirastu R, Casson F, Terracini B, 2006. Environment and health in Sardinia. Epidemiol Prev 30, 1-96.

Carlin BC, Louis TA, 2009. Bayesian Methods for Data Analysis. 3rd edition. CRC/Chapman & Hall, Boca Raton, FL, USA.

Catelan D, Lagazio C, Dreassi E, Pirastu R, Terracini B, Biggeri A, 2006. Statistical approaches to environmental epidemiology of high risk areas. The Sardinia region (Italy) report. Proceedings XXIII International Biometric Conference, Montreal, Canada.

Carpenter LM, Maconochie NES, Roman E, Cox DR, 1997. Examining associations between occupation and health by using routinely collected data. J R Stat Soc Ser A 160, 507-521.

Elliott P, Wakefield J, Best N, Briggs D, 2000. Spatial Epidemiology - Methods and Applications. Oxford University Press, Oxford, UK.

Gilks WR, Richardson S, Spiegelhalter DJ, 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall, London, UK.

Goldstein H, Spiegelhalter DJ, 1996. League tables and their limitations: statistical issues in comparisons of institutional performance, with discussion. J R Stat Soc Ser A 159, 385-443.

Grisotto L, Catelan D, Lagazio C, Biggeri A, 2007. L'uso dell'indice di deprivazione materiale in epidemiologia descrittiva. In: Comba P, Bianchi F, Iavarone I, Pirastu R (Eds). Impatto sulla salute dei siti inquinati: metodi e strumenti per la ricerca e le valutazioni. Rapporti ISTISAN 07/50, Istituto Superiore di Sanità, Roma, Italy, pp. 123-134.

Laird NM, Louis TA, 1989. Empirical Bayes ranking methods. J Educ Stat 14, 29-46.

Law G, Cox DR, Machonochie N, Simpson J, Roman E, Carpenter L, 2001. Large tables. Biostatistics 2, 163-171.

Mitis F, Martuzzi M, Biggeri A, Bertollini R, Terracini B, 2005. Industrial activities in sites at high environmental risk and their impact on the health of the population. Int J Occup Environ Health 11, 88-95.

Rothman K, 2002. Epidemiology: An Introduction. Oxford University Press, New York, USA.

Shen W, Louis TA, 1998. Triple-goal estimates in two-stage hierarchical models. J R Stat Soc Ser B 60, 455-471.

Spiegelhalter DJ, Thomas A, Best N, Lunn D, 2003. WinBUGS 1.4. Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University, Cambridge, UK.

Storey J, 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat 31, 2013-2035.

Vigotti MA, Biggeri A, Dreassi E, 2001. Atlante della Mortalità in Toscana 1971-1994. Edi Plus, University of Pisa, Pisa, Italy.