Bayesian binomial regression with change point prior for the analysis of the risk around a point source

Annibale Biggeri[1,2], Dolores Catelan[1,2], Linda Guarda[3], Corrado Lagazio[4]

[1] Department of Statistics "G.Parenti", University of Florence, Viale Morgagni, 59, 50134 Florence, Italy
[2] Biostatistics Unit, ISPO Cancer Prevention and Research Institute, Via Cosimo il Vecchio, 2, 50139 Florence, Italy
[3] Local Health Authority, ASL Mantua, Italy
[4] Department of Statistical Sciences, University of Udine, Via Treppo 18, 33100 Udine, Italy

Corresponding Author:
Dr. Dolores Catelan
Dept. of Statistics "G. Parenti", viale Morgagni, 59 - 50134 Florence, Italy
Fax +39 055 3269 7950 - Tel: +39 055 3269 7948
e-mail: catelan@ds.unifi.it

Summary

Objective: in recent years there has been a growing interest in the study of risk of disease around a point source of pollution. A number of statistical methods has been proposed to address such analysis with Poisson data. We focus on the analysis of case-control data taking advantage of a real example regarding mortality for respiratory causes around the oil-powered energy plant located in Ostiglia, Mantua (IT).

Methods: Isotonic regression was used as a tool for data description. We proposed a Bayesian Binomial regression model with change-point prior without specifying any constraint on the risk pattern.

Results: We analyzed 109 cases of deaths for respiratory diseases in the period 1995-1998 and 355 controls. Distance from residence to putative source was used as proxy of exposure. An excess (relative risk 1.50; 90% credible interval 1.18; 1.83) of mortality for respiratory causes within 4.5 km from the Ostiglia plant was found.

Conclusion: The Bayesian Binomial regression with change-point prior is a flexible approach to estimate the risk around a putative point source under a case-control design.

KEY WORDS: *Point Source Analysis, Case-Control, Bayesian Binomial Regression, Change-Point Prior, Isotonic Regression*

Introduction

In the last twenty years there has been a growing interest in the study of disease risk around a point source of pollution (1). Such studies are known in the scientific literature as point-source studies (2). To analyze those data a common approach is to consider distance from the source as a proxy of exposure (3); in particular, assuming an isotropic process for the risk, it is possible to consider classes of distance, built up as circular annuli centered on the source (4).

The statistical analysis depends on the type of available information: we may have data only at aggregate level, that is counts of cases and population denominators by area or we may know the exact location of each single case event. Moreover it depends on the knowledge about the population distribution in the study area. If population density is known, then the analysis proceeds comparing observed and expected number of disease cases. Tests based on the difference observed-expected weighted by distance from source (5,6), on the ratio observed/expected by circular annuli of growing radius centred on the source (7) or Poisson probability tests (8) have been proposed in the literature. When population density is not known the spatial distribution of non cases can be estimated using a "control" sample. Indeed, the spatial distribution of cases is not informative per se since population at risk is ge-

nerally not homogeneously distributed.
Poisson or Logistic regression models in which a parametric function of distance from the source is specified are available in the literature (9), while non parametric approaches were widely used only for Poisson data (7, and for an example in case control data 10). For binomial data, non parametric tests which had been applied, are the cumulative and the maximum chi square tests (11,12). These tests are powerful in determining the presence of an association between distance and risk, permit to determine a threshold for the effect but lack in quantifying the risk and estimating the response function.
When the relationship between outcome and exposure is monotonic and we do not want to make strong assumptions on the shape of the response function the isotonic regression is appropriate (13). The only assumption is that the response cannot decrease as exposure increases. This kind of analysis is useful also when we are interested in determining the potential number of steps or breaking points and their position in case the risk function be discontinuous.
In this work we propose a Bayesian approach with change point prior to point source analysis when case-control data are available. A similar formulation has been proposed in the literature for Poisson data (14). We take advantage of the real example of the Ostiglia study (Mantua, Italy), where a oil-powered energy plant is located (15).

Data

The original study
Data come from a case-control study (15) carried out in the year 2000. The study was commissioned by the Province of Mantua (Italy) to screen the health status of the population living near the power plants of Ostiglia and Sermide. The original study area covered 17 municipalities in the south-west part of the Province of Mantua for a total population of 46,549 inhabitants at the moment of the investigation. Five causes of death were considered: ischemic heart disease (ICD-IX: 410-414), cerebrovascular diseases (ICD-IX: 430-438), respiratory diseases (ICD-IX: 460-519), lung cancer (ICD-IX: 162), lymphoematopoietic malignant tumors (ICD-IX: 200-208). All death certificates with underlying cause of death listed before of all residents dead in the study area on the calendar period 1995-1998 were enrolled as cases. For each case up to 4 controls of the same sex and age, resident and present in the same area at death, and dead in the same period of the case (±2 years) for causes not related to the exposure of interest were identified. Mortality data, both for cases and controls, came from the Mortality Register of the Local Health Authority of Mantua (ASL).
Each case and control were assigned to the 1991 census track of residence. For each census track information on a series of variables as level of education, unemployment, house tenure and house crowding were available and were used as confounding variables in the analysis.
The distance between residence at death and putative point sources was used as proxy of exposure.
The original report (15) highlighted an excess death risk for respiratory diseases for residents within 4,5 Km from the Ostiglia power plant (Adjusted Odds Ratio OR*adj* = 2.13; 95% confidence interval CI= 1.16-3.91).

Materials
In the present analysis we restrict our attention to 109 cases of deaths for respiratory diseases (ICD-IX: 460-519) in the period 1995-1998 and 355 controls. We used as a proxy of exposure the distance from Ostiglia power plant categorized into 13 circular annuli (1 km width) centered on the point source. In Table 1 we report the distribution of the number of cases and controls by distance.

Methods
Let assume to have N circular bands centered on a putative point source. For the i-th band let define Y_i, (i=1,…,N) the number of cases, n_i the sum of cases and controls and p_i the probability of being a case.
The vector of probabilities $\mathbf{p}=\{p_1,\ldots, p_i ,\ldots, p_N\}$ represents the risk gradient by distance, assuming constant risk within each band. We further assume that the proportion p_i be a piecewise function of distance and that it may vary k times, with $0 \le k \le N$-1.
Let define s_j the position of the change point (the location in which p_i change) and assume that $1<s_1<s_2<\ldots<s_k<N$ ($s_0=1$ if k=0). By definition the proportion of cases between the locations s_j and s_{j+1} is $p_{s_{j+1}}$. If k=0 the risk is constant, i.e. $p_i=p\ \forall\ i$.
Let assume Y_i be conditionally independent given the parameters p_{s_j}, s_j e k, and distributed as
$Y_i \mid p_{s_j}, s_j, k \sim$Binomial$(p_{s_j}, n_i)$, $i = 1,\ldots, N$.

The joint likelihood is therefore:

$$L(\mathbf{Y} \mid p,s,k,n) = \prod_{i=1}^{S_1}\left\{\binom{n_i}{y_i} p_{s_1}^{\ y_i}(1-p_{s_1})^{n_i-y_i}\right\}\cdot\ldots\cdot\prod_{i=s_k}^{N}\left\{\binom{n_i}{y_i} p_{s_{k+1}}^{\ y_i}(1-p_{s_{k+1}})^{n_i-y_i}\right\}$$

A linear model is specified on the logit of p_i:

$$\text{logit}(p_i) = \beta_1 + \beta_2\, \text{I}(d_i \geq s_1) + \beta_3\, \text{I}(d_i \geq s_2) + \ldots + \beta_{k+1}\, \text{I}(d_i \geq s_k)$$

where I(·) is the indicator function, d_i indicate the circular band, β_1 is the logarithm of the odds up to the first change point, $\beta_1+\beta_2$ from the first to the second change point, $\beta_1+\beta_2+\beta_3$ between second and third and so on till the k-*th* change point.

Bayesian inference requires to specify the prior knowledge. This is achieved by eliciting prior distributions on the unknowns parameters (e.g. the positions s_i, and the β_k risk terms). In our analysis we fixed the number of change points to be small and we specified Uniform a priori distributions in (1, …, *N*-1) on the change points locations s_j (with the order restriction $s_1 < \ldots < s_k$). β parameters are assumed a priori distributed as Normal with 0 mean and small precision.

The chosen change-point prior results in a smoothed risk function.

Posterior distributions are approximated by MCMC algorithm. We have made use of the WinBUGS software (16). Table 2 reports the part of the WinBugs code which referred to the binomial likelihood and the linear predictor. The code for the a priori specifications can be found in Kokki and Penttinen (14).

We also made an explorative analysis by maximum likelihood isotonic regression running the PAVA and CIR.PAVA algorithms of R software (17). That is, we modelled the risk via a non parametric function which have the only constrain of monotonicity of the relation between risk and distance from the point source (13). The CIR PAVA algorithm (Centered Isotonic Regression) is developed to estimate true smooth dose-response functions (18). The PAVA algorithm (Pool Adjacent Violators Algorithm) constrains observations violating the monotonicity assumption to a weighted average of adjacent areas. This leads to piece-wise-constant flat intervals (19).

Results

Estimates of the relative risk (RR) as function of distance obtained by maximum likelihood isotonic regression thorough the PAVA and CIR.PAVA algorithms are reported in Figure 1.

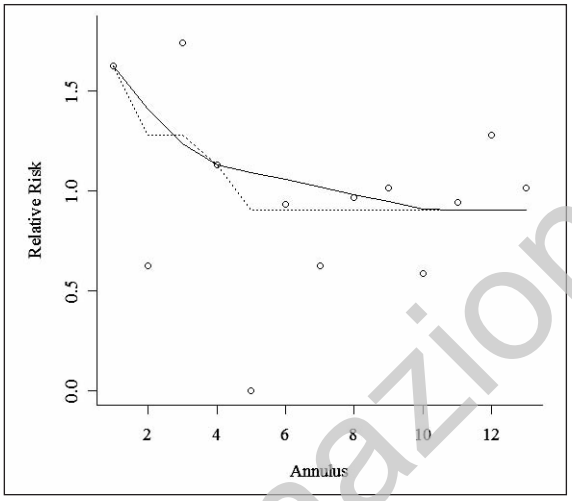

Figure 1. Maximum likelihood isotonic regression. Estimated risk function by distance from residence to the Ostiglia power plant. PAVA (dotted line) and CIR.PAVA Algorithm estimates (solid line) (see text). Mortality for Respiratory causes, Mantua, 1995-1998.

Distance is expressed in km and corresponds to the index number of each band (see Table 1). Relative Risks (RR) are obtained dividing the predicted proportion in each band by the overall proportion of cases in the sample (0.23). The estimated RR in the first band is 1.6 RR, decreases by distance and it is null (RR=1) from the fifth band onward (around 5 Km from the plant). The function seems to suggest the presence of two change

Table 1. Distance from residence to the Ostiglia power plant (in classes), annulus index number, counts of cases and controls in each annulus. Mortality for respiratory diseases, Mantua, 1995-1998.

Annulus	Distance	Cases	Controls	Total
1	0-1,5 km	22	37	59
2	1,5-2,5 km	1	6	7
3	2,5-3,5 km	4	6	10
4	3,5-4,5 km	7	20	27
5	4,5-5,5 km	0	1	1
6	5,5-6,5 km	3	11	14
7	6,5-7,5 km	6	36	42
8	7,5-8,5 km	4	14	18
9	8,5-9,5 km	21	69	90
10	9,5-10,5 km	7	45	52
11	10,5-11,5 km	8	29	37
12	11,5-12,5 km	5	12	17
13	12,5-inf km	21	69	90
	Total	**109**	**355**	**464**

points, one after the second band and one after the fifth. We fixed k=2 number of change points in the Bayesian formulation without constraints on monotonicity.
In table 3 we report parameter estimates from the Bayesian regression model. β coefficients (log Odds) taking as reference the log odds of the overall proportion of cases (log(0.23/0.77)=-1.21) indicates a decreasing behaviour of relative risk by distance from the source. Posterior estimates of change points position are 3.5 and 7.7 km with large 95% CrI (1.1-7.0 for the first, 2.8-12.4 for the second). The excess risk in proximity of the plant vanishes after the seventh band. Figure 2 reports the predicted RR as function of distance with 80% credibility bands (CrI) and the estimated risk function by maximum likelihood isotonic regression (PAVA).

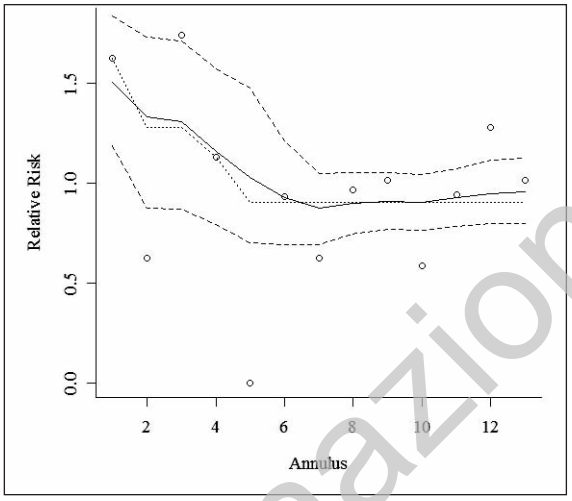

Figure 2. Posterior risk function by distance from residence to the Ostiglia power plant. Bayesian Binomial regression with change-point prior (solid line) with 80% Credibility band (dashes lines) and maximum likelihood isotonic regression (PAVA algorithm, dotted line) (see text). Mortality for Respiratory causes, Mantua, 1995-1998.

Discussion and Conclusions

Our Bayesian analysis of Ostiglia's data provided evidence of an excess mortality risk for respiratory causes within 4,5 Km from the plant. This is consistent with the previous report (15) and with the exploratory analysis. In the epidemiological literature a short term effect of fine particulate on mortality for respiratory diseases has been widely documented (20). The present study design did not permit to disentangle short and long term effects since we used residential address at death. However the magnitude of relative risk found was consistent with the literature on long term effect (22, 23). The approach used suffers of the arbitrariness in the construction of the bands and, in particular, in the choice of their size. However, provided that the dimension of the bands be sufficiently small, that approach is sensible in studying the relation of interest (21).
The Ostiglia case is used only as a motivating example. In this work we addressed case-control design and proposed a Bayesian model with change point prior to analyze disease risk gradient by distances. Closely related statistical methods have been proposed for Poisson data (14) but up to date not for binomial data.
The Bayesian model is easy to implement with standard statistical software, as WinBugs (16). The definition of circular bands around the point source reduces a "spatial problem" to a one-dimensional problem. The choice of circular bands is justified by the isotropy of the process. If the assumption of isotropy is not valid, other subdivision of the space are possible, such as ellipses.

Table 2. Part of the WinBUGS code for the Bayesian Binomial Regression with change-point prior.

```
model
{
for (i in 1:N)
{
n[i]<-y[i]+cn[i]
y[i]~ dbin(p[i],n[i])
logit(p[i])<-beta[1]+step(i-s[1])*beta[2]+step(i-
s[2])*beta[3]
}
for (j in 1:3)  { beta[j]~ dnorm(0,1.0E-06) }
}
```

An important issue is the choice of the number of change points. It may have an impact in the degree of smoothing of the predicted risk function. One can leave the number of change points as a parameter to be estimated in the model but the analysis would then become extremely complex and computationally heavy. Generally speaking, data from point source studies are not informative on the number of change points (14). In the data set currently analyzed the study area was not spread out and fixing a small number of change points seemed a reasonable choice. However, a sensitivity analysis was performed to assess the impact of different number of change points (not shown).
In conclusion, we focused on the analysis of risk gra-

Table 3. Posterior estimates of parameters from the Bayesian Binomial regression with change-point prior (see text). Mortality for Respiratory causes, Mantua, 1995-1998.

Parameter	Posterior mean	Standard deviation	Centiles of posterior distribution				
			2.5%	10%	mediana	90%	97.5%
s[1]	3.5	1.78	1.1		3.4		7.0
s[2]	7.7	2.78	2.8		7.6		12.4
Beta[1]	-0.65	0.26	-1.15		-0.65		-0.12
Beta[2]	-0.69	0.62	-1.99		-0.69		0.57
Beta[3]	0.06	0.62	-1.20		0.08		1.36

dients as function of distance from putative point sources with individual case-control data. This kind of design can be viewed as a sort of mixed design, partly based on individual information and partly based on ecological measurement, distance from residence to source. In this sense it is superior to purely ecological studies based on aggregate data.

On the other side, this design is subject to biases other than those which could affect any case-control study. Theoretically the relevant location should be that at which the exposure could have been experienced. Examples of incorrect location definition are for the cases, the residence at death when analyzing effect with a long latency and large migratory population flows are present, and for the controls, a more recent residence than the index case (24). Examples of imperfect assessment of location are the use of geographical maps at low resolution incomplete recovery of residential histories (25).

The definite merits of this sort of investigation stands on the ability to model the risk gradient while adjusting for relevant individual risk factors in all instances where historical data on exposure are difficult to obtain or unreliable.

The proposed Bayesian Binomial regression with change-point prior can be a useful non parametric alternative to analyze the risk of disease around a point source when case-control data are available.

It is valuable in all cases in which we cannot make strong assumptions on the shape of the response function.

Aknowledgments

Linda Guarda was partially supported by the Master in Epidemiology, University of Turin and San Paolo Foundation.

References

1. Kibble A, Harrison R. Point sources of air pollution *Occupational Medicine* 2005; SS(6):425-431.
2. Elliott P, Wakefield JC, Best NG, Briggs DJ. *Spatial epidemiology: methods and applications* In: Elliott P, Wakefield JC, Best NG, Briggs DJ, eds. Spatial Epidemiology, Oxford University Press, Oxford, 2000.
3. Diggle PJ. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a pre-specified point. *Journal of the Royal Statistical Society A* 1990; 153:340-362.
4. Shaddik G, Elliott P. Use of Stone's Method in Studies of Disease Risk Around Point Sources of Environmental Pollution. *Statistics in Medicine* 1996; 15:1927-1934.
5. Lawson AB. On the analysis of mortality events associated with a pre-specified fixed point. *Journal of the Royal Statistical Society A* 1993; 156:363–377.
6. Tango T. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine* 1995; 14:2323-2334.
7. Stone RA. Investigation of Excess of Environmental Risk around Putative Sources*:* Statistical Problems and a Proposed Test. *Statistics in Medicine* 1988; 7:649-660.
8. Besag J, Newell J. The Detection of Clusters in Rare Diseases. *Journal of the Royal Statistical Society A* 1991; 154(1):143-155.
9. Biggeri A, Lagazio C. *Case-control Analysis of Risk around Putative Sources*. In: Lawson AB, Biggeri A *et al.* (eds) *Disease Mapping and Risk Assessment for Public Health*. Wiley, Chichester, 1999.
10. Cuzick J, Edwards R. *Spatial Clustering for Inhomogeneous Populations (with discussion). Journal of the Royal Statistical Society B* 1990; 52: 73-104.
11. Hirotsu C. Defining the Pattern of Association in Two-Way Contingency Tables. *Biometrika* 1993; 70:579-589.
12. Nair VN. Chi-Squared Type Tests for Ordered Alternatives in Contingency Tables*. Journal of the American Statistical Association* 1987; 82:283-291.
13. Morton-Jones T, Diggle P, Parker L, Dickinson HO, Binks K. Additive isotonic regression models in epidemiology. *Statistics in Medicine* 2000; 19:849-859.
14. Kokki E, Penttinen A. Poisson Regression with Chan-

ge-Point Prior in the Modelling of Disease Risk around a Point Source. *Biometrical Journal*, 2003; 45(6): 689-703.
15. Biggeri A, Dreassi E, Giannella G, Talassi F, Alessi F. *Studio sulla distribuzione geografica dei deceduti per malattie ischemiche di cuore, malattie cerebrovascolari, malattie dell'apparato respiratorio, tumore polmonare, tumori del sistema linfoematopoietico e distanza dagli impianti termoelettrici di Ostiglia e Sermide.* Provincia di Mantova, 2002.
16. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 2000; 10: 325-337.
17. Salanti G. *The isotonic regression framework – Estimating and testing under order restrictions.* PhD Dissertation, University of Padua, 2003.
18. Ramsay JO. Estimating smooth monotone functions. *Journal of the Royal Statistical Society B* 1998; 60: 365-375.
19. Robertson T, Wright FT, Dykstra RL. *Order Restricted Statistical Inference*. Wiley, Chichester, 1988.
20. WHO. *Air quality guidelines. Global update 2005*. WHO Regional Office for Europe, Copenhagen, Denmark, 2006.
21. Lawson AB. *Statistical methods in spatial epidemiology*. Wiley, Chichester, 2001.
22. Abbey DE, Nishino N, McDonnell WF, Burchette RJ, Knutsen SF, Lawrence Beeson W, Yang JX. Long-Term Inhalable Particles and Other Air Pollutants Related to Mortality in Nonsmokers. *American Journal of Respiratory and Critical Care Medicine* 1999; 159:373–382.
23. Brunekreef B, Beelen R, Hoek G, Schouten L, Bausch-Goldbohm S, Fischer P, Armstrong B, Hughes E, Jerrett M, van den Brandt P. *Effects of Long-Term Exposure to Traffic-Related Air Pollution on Respiratory and Cardiovascular Mortality in the Netherlands: The NLCS-AIR Study*. HEI Research Report 139, Health Effects Institute, Boston, MA, 2009.
24. Ross A, Scott D. Point Pattern Analysis of the spatial proximity of residences prior to diagnosis of persons with Hodgkin's disease. *Americam Journal of Epidemiology* 1990; 132:53-62.
25. Diggle PJ, Elliott P. Statistical issues in the analysis of disease risk near point source using individual or spatially aggregated data. *Journal of Epidemiology and Community Health* 1995; 49(Suppl.2): S20-S27.