International Journal of Health Geographics

BioMed Central

Research

Open Access

Robustness of the BYM model in absence of spatial variation in the residuals

Aurélien Latouche*[1,2], Chantal Guihenneuc-Jouyaux[1,2,3,4], Claire Girard[1,2] and Denis Hémon[1,2]

Address: [1]Inserm, U754, Villejuif, F-94807, France, [2]Univ Paris-Sud, IFR69, Villejuif, F-94807, France, [3]Univ Paris 5, Paris, F-75006, France and [4]CNRS, UMR8145, Paris, F-75006, France

Email: Aurélien Latouche* - aurelien.latouche@paris7.jussieu.fr; Chantal Guihenneuc-Jouyaux - guihenneuc@biomedicale.univ-paris5.fr; Claire Girard - girard@vjf.inserm.fr; Denis Hémon - hemon@vjf.inserm.fr

* Corresponding author

Published: 20 September 2007

International Journal of Health Geographics 2007, **6**:39 doi:10.1186/1476-072X-6-39

Received: 31 May 2007
Accepted: 20 September 2007

This article is available from: http://www.ij-healthgeographics.com/content/6/1/39

© 2007 Latouche et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: In the context of ecological studies, the Bayesian hierarchical Poisson model is of prime interest when studying the association between environmental exposure and rare diseases. However, adding spatially structured extra-variability in the model fitted to the data when such extra-variability does not exist conditionally on the covariates included in the model (*over-fitting*) may bias the estimation of the ecological association between covariates and relative risks toward the null. In order to investigate that possibility, a simulation study of the impact of introducing unnecessary residual spatial structure in the estimation model was conducted.

Results: In the case where no underlying extra-variability from the Poisson process exists, the simulation results show that models accounting for structured and unstructured residuals do not underestimate the ecological association, unless covariates have a very strong autocorrelation structure, i.e., 0.98 at 100 km on a territory of diameter 1000 km."

1 Background

Ecological regression studies investigate potential association between geographical variation in disease rates (or counts) and environmental covariates. For example, a recent study evaluated the ecological association between indoor radon concentration and acute leukaemia incidence among children [1]. For rare diseases and/or small areas, Bayesian hierarchical Poisson model is commonly used where within-area variability of disease is modelled at the first stage as a Poisson process and ecological relationships between disease and covariates are introduced at the second stage of the hierarchical model. Spatially extra-Poisson variability potentially due to aggregated effect of unknown confounders is commonly taken into account through spatially structured residuals added in the second stage of the model.

In that context, the BYM (Besag, York and Mollié) model [2] is a standard model for estimation of the ecological associations. The overall variability of a health indicator is broken down into a random Poisson component, a spatially structured area-specific random effect and an unstructured random term, across geographic units. It has been extensively shown that not accounting for an actual spatial variability may lead to major biases [3].

Conversely, if the spatial variability of a health indicator is completely explained by that of environmental factors

and the other ecological covariates taken into consideration, regression residuals do not have spatial structure. Modelling the spatial structure of residuals could then lead to a biased estimate of the ecological association via a phenomenon of *over-fitting* [4-7].

To the author's knowledge, the quantitative impact of fitting a model including extra-Poisson variability to analyse data generated by a model where such extra-Poisson variability does not exist conditionally on the covariates included in the model (*over-fitting*) has not previously been explicitly and quantitatively investigated. Robustness of residuals modelling as BYM was studied in a different inferential context. In the frame of an extensive investigation of the statistical performances of a number of spatial models, Lawson *et al*. [8] studied the performance of such models on relative risks estimates in the case of mapping modelling, i.e. without covariates, where different true spatial structures of residuals were simulated. The authors showed that BYM model performed well on risks estimations except when the true residuals were resulting from a mixture structure. Different models were compared to detect effect from a putative source on a regular lattice [9]. They concluded in particular that the introduction of a spatially structured area-specific random effect leads to much less bias in the parameter estimate and that BYM model is the least biased. Notably, biases in parameter estimation appear when the random effects are not acounted for. In the present study we focused on ecological association estimate when covariates with spatially structure are introduced. Such covariates are often of interest in epidemiology when environmental exposures are studied. In our work we focus on a particular model with France mainland as study domain. The aim of our study was then to determine the robustness of the BYM model in the absence of residual spatial variation, i.e., the impact on the ecological association estimate on an irregular domain. The estimates performances were discussed and characterized according to the covariates structure. Simulations protocols assumed systematically a Poisson model at the first stage of the hierarchical model and log linear relationship between the incidence of the disease and exposure at the second stage without addition of extra-Poisson residuals. Various spatial structures of exposure were considered. The explained spatial variability is thus fully specified/attributable by the covariate structure.

First, ecological models that do or do not allow spatially structured or unstructured heterogeneity will be considered. Various simulation protocols for parameter values enabling balanced or unbalanced between/within area variability will then be presented. The results of the various simulation protocols will then be considered in terms of their performances with regard to the estimation of ecological associations. The paper will conclude with a discussion.

2 Methods

2.1 Statistical Models

Let D be the study area of interest, partitioned into m geographic areas. The data consist in the observed Y_i and expected E_i disease counts for each area i, $(i = 1,...,m)$. Let X_i be an ecological variable of interest in area i. The ecological Poisson M_0 model is expressed in hierarchical form as follow:

$$\begin{array}{lcl} Y_i \mid R_i & \sim & \text{Poisson}\,(E_i R_i) \\ \log(R_i) \mid \alpha, \beta, X_i & = & \alpha + \beta X_i \end{array} \qquad (M_0)$$

in which R_i is the relative risk in area i. The second stage models the relationship between the relative risk and exposure variable. The ecological model M_0 does not include any spatially structured or unstructured heterogeneity.

In order to account for those variabilities, the BYM model [2] was proposed :

$$\begin{array}{lcl} Y_i \mid R_i & \sim & \text{Poisson}(E_i R_i) \\ \log(R_i) \mid \alpha, \beta, U_i, V_i & = & \alpha + \beta X_i + U_i + V_i \\ U_i \mid U_{j \neq i} & \sim & N(\sum_{j \in \delta_i} U_j / n_i, 1/\tau_U^2 n_i) \\ V_i \mid \tau_V & \sim & N(0, \tau_V^2) \end{array} \qquad (BYM)$$

in which δ_i denotes the set of labels of the neighbours of area i, n_i is the number of neighbours i, U_i $(i = 1,...,m)$ models the spatially-structured area-specific random effect based on the conditional autoregressive approach CAR [10], and V_i $(i = 1,...,m)$ is the unstructured random effect. The BYM model is the benchmark parametric model and is widely used in disease-mapping studies mainly because of the flexibility of the residuals.

2.2 Design of the simulation study

Processes X and Y are simulated on D under model M_0 accordingly to parameters values. The simulation parameters were selected with reference to the overall variability of the estimated relative risks, thus enabling realistic and reasonable values for relative risks. More precisely, let $\hat{R}_i = Y_i/E_i$ be the maximum likelihood estimate of the relative risk for the area. If $X_i \sim N(0, 1)$, then:

$$Var\left[\log(R_i)\right] \approx E\left(\frac{1}{E_i R_i}\right) + \beta^2.$$

If the relative risks are spatially independent of the expected disease counts E_i,

$$E(\frac{1}{E_i R_i}) \approx 1/\bar{E}_h \times E(1/R_i),$$

where $\bar{E}_h$ is the harmonic mean of E_i (i = 1,...,m). The overall variance may be expressed as:

$$Var\left[\log(R_i)\right] \approx 1/\bar{E}_h \times \exp(-\alpha + \beta^2) + \beta^2.$$

This variance may be broken down into a *Within area variability* term Wv = $1/\bar{E}_h$ × exp(-α + β^2) and a *Between area variability* term Bv = β^2. Let p denote the proportion of between area variance, p = Bv/*Var*[log($\hat{R}_i$)], a high value of p corresponds to high between-area variability, that is a high amount of information with which to estimate the ecological link β. Hereafter, without any loss of generality, α will be considered equal to 0.

The geographic scale unit consisted in the 94 *Departements* of mainland France (Corsica excluded). The expected disease counts (E_i) consist in the expected cases of acute leukaemia in children aged less than 15 years for the period 1990–1998 *in Departement i*. The cases were retrieved from the French National Registry of Childhood Leukaemia and Lymphoma [11]. The expected numbers ranged from 4.2 to 204, with a harmonic mean of 23.35. Scenarios in which the within-area variance was either doubled ($\bar{E}_h$ = 46.6) or divided by 10 ($\bar{E}_h$ = 2.33) were also considered.

Given that $X_i \sim N(0, 1)$, within-area variance depends on 3 parameters, namely: the harmonic mean of expected disease counts $\bar{E}_h$, the ecological link β and the autocorrelation structure of X_i.

As X has a standardized normal distribution, the 2.5 ($p_{2.5\%}$) and 97.5 ($p_{97.5\%}$) percentiles of the relative risks are under model M_0 exp($\alpha \pm 1.96\beta$) and their ratio $Q = p_{97.5\%}/p_{2.5\%}$ is exp(2 × β × 1.96). The quantile ratios were considered equal to 1.0, 1.5, 2.0 and 3.0, equivalent to no effect, weak, moderate and strong effects, respectively. The corresponding values for β were 0.00, 0.12, 0.21 and 0.33. The proportions of between-area variance, p, by ecological link β and $\bar{E}_h$ breakdown are summarized in Table 1.

For $\bar{E}_h$ = 23.35, the between area variance proportion ranges from 0 to 71%, with a balanced case for an ecological link when β = 0.21. For $\bar{E}_h$ = 46.69, p ranges from 0 to 83%, while p ranges from 0 to 19% when $\bar{E}_h$ = 2.33.

The autocorrelation of the exposure variable was also modulated. The following exponential autocorrelation structure was considered: cov(X_i, X_j) = $exp(-d(i, j)\phi)$, in which $d(i, j)$ is the distance between areas i and j. Let ρ_{xx} = $exp(-100\phi)$ be the autocorrelation of two areas 100 km distant from each other. The following values for ρ_{xx} = (0.40, 0.90,0.95, 0.98) were studied. That correlation structure is shown in Figures 1. High values of ρ_{xx} may mimic a spatial bloc structure.

For each combination of parameters ($\bar{E}_h$, β, ρ_{xx}), 400 replicates of (X, Y) = ((X_i, Y_i), i = 1,...,N) were generated using the M_0 model. For each replication, the ecological link was estimated by both models (M_0 and BYM) in a Bayesian framework.

The estimations were made with BRugs [12] software. For each data set, a burn-in of 5000 iterations was used and Bayesian inferences were based on 45000 iterations from

Table 1: Between area variance proportion, p[3], according to ecological link β and $\bar{E}_h$

β[1] \ $\bar{E}_h$[2]	2.33	23.35	46.6
0.00	0.00	0.00	0.00
0.12	0.03	0.25	0.40
0.21	0.09	0.50	0.67
0.33	0.19	0.71	0.83

[1] β: Ecological link

[2] $\bar{E}_h$: Harmonic mean of expected disease counts

[3] p = Bv/*Var*[log($\hat{R}_i$)]: Between area variance proportion

[*] : Indicator of p-value < 5% from McNemar's test (1)(2)(3) for comparing proportions 1 - π(0) under M_0 and *BYM* models.

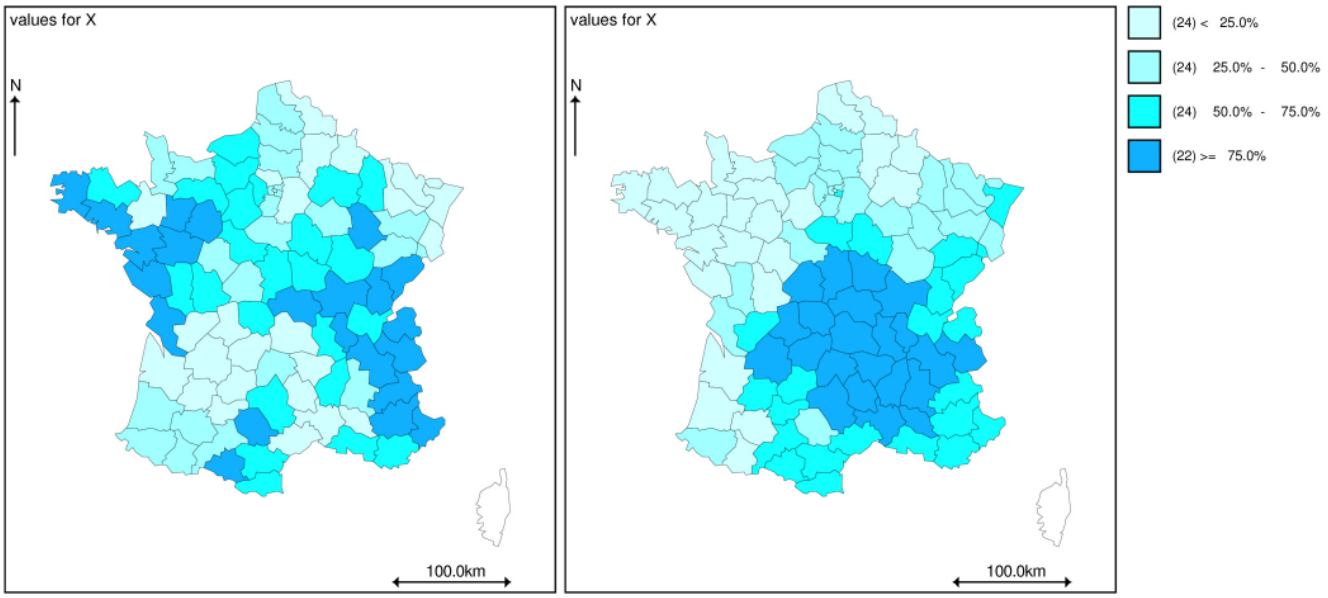

Figure 1
Replicates of a gaussian covariate with autocorrelation strenght of ρ_{xx} = 0.80 (left) and ρ_{xx} = 0.98 (right) at 100 km, (in quartiles).

Gibbs sampling giving Monte Carlo standard errors of less than 5% of the posterior standard deviation of each parameter [13]. The Monte Carlo standard error is an estimate of the difference between the mean of the sampled values and the true posterior mean. Non-informative priors were chosen for the parameters: $\alpha \sim U(-\infty; +\infty)$, $\beta \sim N(0.0, 1.0E+5)$, $\tau_U \sim \Gamma(0.5, 0.0005)$, $\tau_V \sim \Gamma(0.5, 0.0005)$ [14], in which $\Gamma(a, b)$ denotes the Gamma distribution with expectation equal to a/b.

For each triplet ($\bar{E}_h$, β, ρ_{xx}), the ecological link was estimated using the M_0 and BYM models. Let $\hat{\beta}^{(j)}$ be the posterior mean estimates of β at the jth replication, $\sigma_{\hat{\beta}^{(j)}}$ the posterior standard error of $\hat{\beta}^{(j)}$ and $CI^j(\beta)$ the 95% credibility interval of β. The following criteria were computed for 400 replications (j = 1,...,400):

• The empirical mean of the estimated posterior expectations of $\beta\left(\bar{\beta}_{sim} = \sum_{j=1}^{400} \hat{\beta}^{(j)} / 400\right)$

• The empirical mean of the estimated posterior standard deviations of $\beta\left(\bar{\sigma}_{\beta_{sim}} = \sum_{j=1}^{400} \sigma_{\hat{\beta}^{(j)}} / 400\right)$

• The empirical standard deviation of the 400 estimated posterior means of β ($sd(\beta_{sim})$)

• The mean relative bias (MB) and its standard deviation (sd(MB)),

• The root mean square error (RMSE),

• The proportion of coverage, $\pi(\beta)$: the percentage of time when β lay within its 95% credibility interval

• The proportion of non-coverage 1 - $\pi(0)$: the percentage of time when 0 did not lie within its 95% credibility interval

When $\beta \neq 0$, 1 - $\pi(0)$ quantifies the ability of the estimation model to detect the existence of an association, which is analogous to the frequentist power. McNemar's test for comparing proportions from paired data (estimates of β from M_0 and BYM based on the same replicated data set) was used to test whether the $\pi(\beta)$ (or 1 - $\pi(0)$) values were significantly different. The over-fit of the BYM model (compared to the M_0 model) was assessed via that criterion.

3 Simulation Results

Simulations results are structured as follow: firstly, we present results for an harmonic mean $\bar{E}_h$ = 23.35 of expected disease counts equal to those from acute leukae-

mia in children for 1990–1998 in France and a covariate *X* with null to moderate autocorrelation. Secondly, for the same harmonic mean $\overline{E}_h$, we study the influence of strong autocorrelations for the covariate X and finally, variations of $\overline{E}_h$ (smaller and greater than 23.35) are explored.

3.1 Moderate covariate autocorrelation

The first scenario considered $\overline{E}_h$ = 23.35 and the autocorrelation equal to 0.0 or 0.4; the results are shown in Table 2. In that setting, the between-area variance varied from 0 to 71%. In the absence of any spatial structure for *X* (ρ_{xx} = 0), the estimate of β was unbiased, irrespective of the estimation model. The mean bias was less than 1% and the RMSE was less than 0.02 for the four values of the ecological link (for both models). The coverage proportions $\pi(\beta)$ were similar for the two models and greater than 94.5%. The coverage proportion of the BYM model was consistently slightly greater than that of the M_0 model. For β = 0, $\pi(\beta)$ was close to 95% and, equivalently, 1 - π(0) was close to 5%. The BYM model thus handles the scenario in which the covariate has no spatial structure.

When the autocorrelation was increased to 0.4, the results were similar to those with the previous setting. The mean bias was less than 1%. There was a slight increase in the RMSE but it remained less than 0.02. The β coverage proportion with the BYM model was greater than that with the M_0 model. The non-coverage proportion was equal to 1, except when there was no association (β = 0). The non-coverage proportions were significantly different when β = 0. The proportion the closest to 5% was obtained with the M_0 model. Irrespective of the value of β, the variability of β was always slightly over-estimated with the BYM model. The coverage proportion of the M_0 model varied from 94.5 to 95.8% (96.8 to 97.5% for the BYM model). In the absence of, or with moderate, autocorrelation, the *over-fitting* effect was not observed. Both models provided an almost unbiased estimate of the ecological link.

3.2 Strong covariate autocorrelation

The scenario of strong autocorrelation for X was then considered: ρ_{xx} = 0.90, 0.95, 0.98 at 100 km with $\overline{E}_h$ = 23.35. The results are shown in Table 3. When ρ_{xx} = 0.90, the mean bias and the RMSE were low (0.03). The mean bias decreased as the value of the ecological link increased,

Table 2: Estimation of the ecological link β when $\overline{E}_h$ = 23.35 and ρ_{xx} = 0.0, 0.4 (400 replications)

ρ_{xx}(1)	β	Model	$\overline{\beta}_{sim}$(2)	$\overline{\sigma}_{\beta_{sim}}$	MB(4)	sd(MB)	sd(β_{sim})(5)	RMSE(6)	$\pi(\beta)$(7)	1 - π(0)(8)
0.0	0.00	M_0	0.000	1.63			1.73	1.73	0.953	0.047
		BYM	0.000	1.76			1.74	1.73	0.958	0.042
	0.12	M_0	0.119	1.61	-0.44	0.69	1.67	1.67	0.930	1.000
		BYM	0.120	1.72	-0.31	0.70	1.68	1.68	0.955	1.000
	0.21	M_0	0.210	1.60	0.49	0.39	1.64	1.64	0.948	1.000
		BYM	0.210	1.72	0.58	0.39	1.66	1.66	0.960	1.000
	0.33	M_0	0.329	1.57	-0.28	0.23	1.58	1.58	0.953	1.000
		BYM	0.329	1.70	-0.24	0.24	1.58	1.58	0.968	1.000
0.4	0.00	M_0	-0.001	1.72			1.76	1.76	0.955	0.045*
		BYM	-0.001	1.90			1.78	1.78	0.975	0.025
	0.12	M_0	0.121	1.73	0.79	0.73	1.75	1.75	0.945	1.000
		BYM	0.121	1.91	0.78	0.74	1.78	1.78	0.968	1.000
	0.21	M_0	0.211	1.72	0.41	0.39	1.66	1.66	0.958	1.000
		BYM	0.211	1.90	0.50	0.39	1.67	1.67	0.973	1.000
	0.33	M_0	0.332	1.70	0.71	0.25	1.68	1.69	0.945	1.000
		BYM	0.332	1.89	0.70	0.25	1.70	1.71	0.973	1.000

(1) ρ_{xx} : autocorrelation at 100 km

(2) $\overline{\beta}_{sim}$: 100 * mean of posterior means

(3) $\overline{\sigma}_{\beta_{sim}}$: 100 * mean of posterior standard deviations

(4) MB : 100 * Mean Bias

(5) sd(β_{sim}) : 100 * standard deviation of posterior means

(6) RMSE : 100 * Root Mean Square Error

(7) $\pi(\beta)$: Coverage proportion $\beta \in CI(\beta)$

(8) 1 - π(0) : Non-Coverage proportion $0 \notin CI(\beta)$

(*) : Indicator of p-value < 5% from McNemar's test (1)(2)(3) for comparing proportions 1 - π(0) under M_0 and *BYM* models.

Table 3: Estimation of the ecological link β when $\overline{E}_h$ = 23.35 and ρ_{xx} = 0.90, 0.95, 0.98

			(1)(2)(3)(4)(5)(6)(7)(8)(*)							
ρ_{xx}(1)	β	Model	$\overline{\beta}_{sim}$(2)	$\overline{\sigma}_{\beta_{sim}}$	MB(4)	sd(MB)	sd(β_{sim})(5)	RMSE(6)	$\pi(\beta)$(7)	1 - $\pi(0)$(8)
0.90	0.00	M_0	0.003	2.96			2.98	2.99	0.945	0.055*
		BYM	0.003	3.39			3.03	3.04	0.970	0.030
	0.12	M_0	0.122	2.98	2.07	1.24	2.99	3.00	0.958	0.973*
		BYM	0.123	3.44	2.51	1.27	3.07	3.08	0.973	0.938
	0.21	M_0	0.213	2.96	1.38	0.79	3.32	3.34	0.930	0.998
		BYM	0.213	3.41	1.31	0.81	3.41	3.42	0.955	0.998
	0.33	M_0	0.331	2.98	0.37	0.47	3.11	3.11	0.943	1.000
		BYM	0.331	3.43	0.39	0.47	3.13	3.13	0.965	1.000
0.95	0.00	M_0	0.002	4.07			4.04	4.04	0.950	0.050*
		BYM	0.002	4.72			4.14	4.14	0.975	0.025
	0.12	M_0	0.121	4.05	1.03	1.82	4.38	4.37	0.945	0.843*
		BYM	0.121	4.70	1.18	1.89	4.54	4.54	0.960	0.775
	0.21	M_0	0.208	3.99	-1.07	0.99	4.18	4.18	0.960	0.988
		BYM	0.208	4.62	-0.80	1.03	4.31	4.31	0.978	0.983
	0.33	M_0	0.333	4.16	0.78	0.66	4.33	4.33	0.963	1.000
		BYM	0.333	4.80	0.98	0.66	4.34	4.35	0.980	1.000
0.98	0.00	M_0	0.002	6.29			6.48	6.48	0.950	0.050*
		BYM	0.002	7.29			6.53	6.53	0.970	0.030
	0.12	M_0	0.121	6.23	1.22	2.49	5.98	5.98	0.953	0.545*
		BYM	0.122	7.20	1.26	2.47	5.99	5.98	0.975	0.420
	0.21	M_0	0.207	6.35	-1.36	1.57	6.61	6.60	0.960	0.845*
		BYM	0.207	7.35	-1.39	1.59	6.67	6.67	0.985	0.782
	0.33	M_0	0.330	6.27	0.09	1.00	6.64	6.63	0.935	0.985
		BYM	0.331	7.28	0.24	1.01	6.69	6.68	0.975	0.978

(1) ρ_{xx} : autocorrelation at 100 km

(2) $\overline{\beta}_{sim}$: 100 * mean of posterior means

(3) $\overline{\sigma}_{\beta_{sim}}$: 100 * mean of posterior standard deviations

(4) MB : 100 * Mean Bias

(5) sd(β_{sim}) : 100 * standard deviation of posterior means

(6) RMSE : 100 * Root Mean Square Error

(7) $\pi(\beta)$: Coverage proportion $\beta \in CI(\beta)$

(8) 1 - $\pi(0)$: Non-Coverage proportion $0 \notin CI(\beta)$

(*) : Indicator of p-value < 5% from McNemar's test (1)(2)(3) for comparing proportions 1 - $\pi(0)$ under M_0 and *BYM* models.

reflecting an increase in between-area variability. The coverage proportion with the BYM model was higher than the coverage proportion with the M_0 model for all values of β. The non-coverage proportions were significantly different for β = 0.00 and β = 0.12, in favor of the M_0 model. For ρ_{xx} = 0.95, the bias was still small and the RMSE increased to 0.04. The non-coverage proportions were again significantly different in the cases in which β = 0.00 and β = 0.12 in favor of the M_0 model. Lastly, for ρ_{xx} = 0.98 and for the first three values of β (β = 0.00, 0.12, 0.21), the non-coverage proportions were significantly different, again in favor of M_0. When the ecological link was null, the non-coverage proportion was again smaller with the BYM model, a consequence of the over-estimation of parameter variability. The β coverage proportions were lower for M_0 for 4 values of β. When the autocorrelation increased from 0.90 to 0.98, the RMSE increased from 2.99 to 6.48, mainly due to the decrease in independent information. A slight increase was observed for all the other criteria. The bias was weak, resulting in very small RMSE and sd(β_{sim}) in both models. At high autocorrelation values, the overall variability of the estimates increased. There was more variability for each β value with the BYM model. This is exemplified by the mean posterior standard deviation, which increased four-fold (for all β values) between the first (ρ_{xx} = 0.00) and last (ρ_{xx} = 0.98) autocorrelation scenario. This was also the case for sd(β_{sim}).

3.3 Variation of (harmonic mean of) expected counts

The next scenario consisted in strong autocorrelation of ρ_{xx} = 0.95 at 100 km with variation in number of expected disease counts. The results for that scenario are shown in Table 4.

For $\overline{E}_h$ = 46.6 (and ρ_{xx} = 0.95), the mean bias and RMSE were smaller than in the scenario in which $\overline{E}_h$ = 23.3. The β coverage rate was greater than 94.5% for both models and the proportion was higher for the BYM model. The non-coverage proportions were significantly different, in favor of the M_0 model, for β = 0.12. For β = 0, the non-coverage proportion was again smaller with the BYM model.

For $\overline{E}_h$ = 2.33, the bias increased (up to 6%) and the RMSE was the highest observed in the various cases (14% approx.). The coverage proportion was smaller than in the previous scenario but greater than 92%. The coverage proportion $\pi(\beta)$ was higher with the BYM model than with the M_0 model. While the non-coverage proportions of 0 were close to 1 (except for β = 0), the proportions decreased by 16% for M_0 and 13% for BYM for β = 0.12. Moreover the non-coverage proportions of 0 for the two models were significantly different and in favor of the M_0 model for β = 0.12, 0.21, 0.33.

Table 4: Estimation of the ecological link β when ρ_{xx} = 0.95 while $\overline{E}_h$ varying

			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\overline{E}_h$	β	Model	$\overline{\beta}_{sim}$[1]	$\overline{\sigma}_{\beta_{sim}}$	MB[3]	sd(MB)	sd(β_{sim})[4]	RMSE[5]	$\pi(\beta)$[6]	1 - $\pi(0)$[7]
46.6	0.00	M_0	-0.001	2.84			2.76	2.76	0.958	0.042*
		BYM	-0.001	3.40			2.85	2.85	0.975	0.025
	0.12	M_0	0.118	2.81	-1.34	1.21	2.91	2.91	0.945	0.960*
		BYM	0.119	3.38	-1.18	1.21	2.91	2.91	0.980	0.930
	0.21	M_0	0.212	2.87	0.94	0.69	2.89	2.89	0.955	1.000
		BYM	0.212	3.47	0.75	0.70	2.95	2.96	0.983	1.000
	0.33	M_0	0.329	2.88	-0.21	0.47	3.10	3.10	0.953	1.000
		BYM	0.330	3.45	-0.08	0.48	3.19	3.19	0.980	1.000
23.3	0.00	M_0	0.002	4.07			4.04	4.04	0.950	0.050*
		BYM	0.002	4.72			4.14	4.14	0.975	0.025
	0.12	M_0	0.121	4.05	1.03	1.82	4.38	4.37	0.945	0.843*
		BYM	0.121	4.70	1.18	1.89	4.54	4.54	0.960	0.775
	0.21	M_0	0.208	3.99	-1.07	0.99	4.18	4.18	0.960	0.988
		BYM	0.208	4.62	-0.80	1.03	4.31	4.31	0.978	0.983
	0.33	M_0	0.333	4.16	0.78	0.66	4.33	4.34	0.963	1.000
		BYM	0.333	4.79	0.98	0.66	4.34	4.35	0.980	1.000
2.33	0.00	M_0	0.007	12.4			13.5	13.4	0.935	0.065
		BYM	0.008	13.3			13.4	13.4	0.948	0.052
	0.12	M_0	0.112	12.5	-6.94	5.38	12.9	12.9	0.953	0.162*
		BYM	0.113	13.4	-5.65	5.43	13.0	13.0	0.965	0.130
	0.21	M_0	0.206	12.8	-1.92	3.45	14.5	14.5	0.927	0.410*
		BYM	0.206	13.8	-1.98	3.46	14.6	14.5	0.940	0.368
	0.33	M_0	0.341	13.0	3.38	2.12	14.00	14.0	0.938	0.738*
		BYM	0.343	13.9	4.05	2.14	14.1	14.2	0.960	0.715

[1] $\overline{\beta}_{sim}$: 100 * mean of posterior means

[2] $\overline{\sigma}_{\beta_{sim}}$: 100 * mean of posterior standard deviations

[3] MB : 100 * Mean Bias

[4] sd(β_{sim}) : 100 * standard deviation of posterior means

[5] RMSE : 100 * Root Mean Square Error

[6] $\pi(\beta)$: Coverage proportion $\beta \in CI(\beta)$

[7] 1 - $\pi(0)$: Non-Coverage proportion $0 \notin CI(\beta)$

[*] : Indicator of p-value < 5% from McNemar's test for comparing proportions 1 - $\pi(0)$ under M_0 and *BYM* models.

High autocorrelations thus appear to influence the *overfitting* effect of the BYM model. The expected disease counts, also modulates the overall accuracy of the estimation. In fact, with highly correlated spatial structure and low disease counts, the bias of the β estimate generated by the BYM increases. But, even with a highly autocorrelated covariate and adequate disease counts, when E_i is doubled, the BYM model estimates the ecological link with little bias.

4 Discussion

A simulation study was conducted in order to assess estimation performance with respect to the ecological association between covariates and health indicators. Key parameters, such as the ecological link, expected disease counts and autocorrelation strength were selected to ensure that the simulation covered realistic situations. The choice of parameters enabled coverage of balanced and unbalanced between- and within-area variabilities.

For moderate autocorrelation structures, both the Poisson model and the BYM model performed well and the estimation performances were similar. Underestimation of ecological links was only observed for high autocorrelations. Overall, the posterior standard deviation of β was slightly over-estimated with the BYM model, resulting in conservative results when the true value of β was null.

The expected disease counts are also of interest because, with a high autocorrelation, the underestimation of the BYM model is present. In practice, this worst-case scenario can nonetheless be found. Except for the extreme scenario, strong spatial structure and low disease counts, both models perform well, even with strong spatial structure. As a consequence, the BYM model can be used to estimate ecological associations without fearing underestimation. The simulation results show that models accounting for structured and unstructured residuals do not underestimate materially the ecological association. The rational is the following: not accounting for an actual spatial variability leads to strong bias. Thus from a practical point of view, the BYM model should be preferred to the Poisson if spatial autocorrelation of covariate is suspected. Moreover, autocorrelation structure will be first investigated via Moran's I test [15].

Acknowledgements

This work was supported by grants AFSSE (RD2004004) and INSERM-ATC (A03150LS). The first author was supported by INSERM (*poste Blanc*)

References

1. Evrard A, Hémon D, Billon S, Laurier D, Jougla E, Tirmarche M, Clavel J: **Ecological association between indoor radon concentration and childhood leukaemia incidence in France, 1990–1998.** *Eur J Cancer Prev* 2005, **14(2):**147-57.
2. Besag J, York Y, Mollié A: **Bayesian Image Restoration With Two applications In Spatial Statistics.** *Ann Inst Statist Math* 1991, **43:**1-59.
3. Salway R: **Statistical Issues in the Analysis of Ecological Studies.** In *PhD thesis* Imperial College School of Medicine; 2003.
4. Clayton D, L B, C M: **Spatial Correlation in Ecological Analysis.** *International Journal of Epidemiology* 1993, **22:**1193-1202.
5. Best NG, Arnold RA, Thomas A, Waller LA, Conlon EM: **Bayesian models for spatially correlated disease and exposure data.** In *Bayesian Statistics 6* Oxford University Press; 1999:131-56.
6. Richardson S: **Spatial models in epidemiological applications.** In *Highly Structured Stochastic Systems, Volume 27 of Oxford Statistical Science Series* Edited by: Green PJ, Hjort NL, Richardson S. Oxford: Oxford University Press; 2003:237-259.
7. Wakefield J: **Sensitivity analyses for ecological regression.** *Biometrics* 2003, **59:**9-17.
8. Lawson AB, Biggeri AB, Boehning D, Lesaffre E, Viel JF, Clark A, Schlattmann P, Divino F: **Disease mapping models: an empirical evaluation. Disease Mapping Collaborative Group.** *Statistics in Medicine* 2000, **19(17–18):**2217-2241. [Comparative Study]
9. Ma B, Lawson AB, Liu Y: **Evaluation of Bayesian models for focused clustering in health data.** *Environmetrics* 2007. [DOI: 10.1002/env.850]
10. Besag J: **Spatial interaction and the statistical analysis of lattice systems (with discussion).** *Journal of Royal Statistical Society, Series B* 1974, **36:**192-236.
11. Clavel J, Goubin A, Auclerc MF, Auvrignon A, Waterkeyn C, Patte C, Baruchel A, Leverger G, Nelken B, Philippe N, Sommelet D, Vilmer E, Bellec S, Perrillat-Menegaux F, Hemon D: **Incidence of childhood leukaemia and non-Hodgkin's lymphoma in France: National Registry of Childhood Leukaemia and Lymphoma, 1990–1999.** *Eur J Cancer Prev* 2004, **13(2):**97-103.
12. Thomas A, O'Hara B: *BRugs: OpenBUGS and its R interface BRugs* 2005.
13. Roberts G: **Markov chain concepts related to sampling algorithms.** In *Markov chain Monte Carlo in Practice* Edited by: Gilks W, Richardson S, Spiegelhalter D. Chapman and Hall; 1996.
14. Kelsall JE: **Discussion of Bayesian models for spatially correlated disease and exposure data.** In *Bayesian Statistics 6* Oxford University Press; 1999:131-56.
15. Moran P: **Notes on continuous stochastic phenomena.** *Biometrika* 1950, **37:**17-23.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral