Statistics in Medicine

An autoregressive approach to spatio-temporal disease mapping

Journal:	*Statistics in Medicine*
Manuscript ID:	SIM-07-0087
Wiley - Manuscript type:	Paper
Date Submitted by the Author:	16-Feb-2007
Complete List of Authors:	Martinez-Beneito, Miguel-Angel; Dirección General de Salud Publica, Area de Epidemiologia López-Quílez, Antonio; Universitat de València, Estadística i I. O. Botella-Rocamora, Paloma; Universidad Cardenal Herrera-CEU, Departamento de Ciencias Físicas, Matemáticas y Computación
Keywords:	bayesian statistics , Markov random fields, spatial statistics

An autoregressive approach to spatio-temporal disease mapping

M. A. Martínez-Beneito[1]*, A. López-Quilez[2], P. Botella-Rocamora[3]

[1] *Área de Epidemiología. Dirección General de Salud Pública. Generalitat Valenciana.*

[2] *Departament D'Estadística e Investigació Operativa. Universitat de Valencia.*

[3] *Departamento de Ciencias Físicas, Matemáticas y Computación. Universidad Cardenal Herrera-CEU.*

SUMMARY

Disease mapping has been a very active research field during the last years. Nevertheless, time trends in risks have been ignored in most of these studies and they can provide information with a very high epidemiological value. Lately, several spatio-temporal models have been proposed: either based on a parametric description of time trends, on independent risk estimates for every period, or on the definition of the joint covariance matrix for all the periods as a Kronecker product of matrices.The following paper offers an autoregressive approach to spatio-temporal disease mapping fusing ideas from autoregressive time series, in order to link information in time; and spatial modelling, to link information in space. It is shown that our approach generalises the Kronecker product proposal. As a result, risk estimates are obtained for every region related to those in their neighbours and to those in the same region at neighbouring periods.

*Correspondence to: Miguel Ángel Martínez Beneito. Área de Epidemiología. Dirección General de Salud Pública. Generalitat Valenciana. C/ Micer Mascó, 31. 46015-Valencia. Spain

1. Introduction

Statistical techniques for disease mapping have become very popular in epidemiology during the last decade. Indeed, several monographs which broadly cover this topic have appeared recently [1, 2, 3, 4]. These methods enable to smooth ecological health indicators, as for example Standardized Mortality Ratios [5], life expectancy [6] or fertility schedules [7]. They also allow to perform ecological regression studies [8] or even survival analysis [9] accounting for the geographical structure of the administrative units under study, and therefore the dependence of their observations. As a consequence, the estimates in less populated areas are more reliable due to the sharing of information between neighbouring regions, which are intended to share common risk factors. Thus, it becomes possible to display the geographical distribution of risk even in "small" areas, in which no statistical treatment would imply greater variance in their estimates. In that case, the most extreme risks would relapse systematically into these regions, distorting the meaning of the corresponding map [10].

Among all the proposals to perform risk smoothing which have appeared in the literature, the one stated by [11] (*BYM* from now on) has had a particular impact, and many applications of such model can be found. This approach decomposes the risk in every region as the sum of two effects. The first one, spatially-dependent, models those factors with inter-regional scope, in order to take into account risk determinants that exceed the limits of one or more geographical units. The second one, independent for all the different geographical units, allows for the existence of neighbouring regions with very distinct behaviour opposing to the spatial term. However, although the *BYM* model is a benchmark in geographical risk estimation, it ignores the temporal evolution of risks in the region under study. Therefore, risk estimates provided with such proposal are supposed to be static in time and this assumption is not always

too realistic, mainly in those problems with a wide observed time window. Nevertheless, despite the interest of spatio-temporal monitoring of disease, the number of applications published on this field have been much lower than for the spatial one, maybe due to the difficulties of linking both the spatial and temporal dependence in a single model.

In the bayesian literature, [12] propose to model spatio-temporal data as an ecological regression model with spatially structured errors in the form of gaussian conditionally autoregressive distributions (CAR from now on), both in the intercept and the covariate, which in this case is the time period. In this approach, risk evolution in time for every region depends on the evolution of its neighbours, but it only allows linear time trends. Following this approach, [13] introduce a quadratic term in time trend but they also rely on a parametric description of such evolution, which they consider to be appropriate for short time periods. In [14], an ecological regression on time with a proper gaussian CAR effect both in the intercept and the covariate is proposed, instead of the improper CAR proposal of [12].

Working also in the bayesian frame, but introducing a very different approach, [15] and [16] model risks for every period as time-independent spatial and heterogeneous effects sharing their precisions for the different time periods. Following this approach [17] structure the distribution of the former precisions as a gaussian random walk in time, so variability in spatial dependence across time is allowed in such a way that those years which are closer have a similar behaviour. Recently, [18] have also proposed a temporally-independent multivariate model, in which lung cancer mortality is jointly studied for both males and females considering possible temporal and spatial correlations between them. On the other hand [19] also introduced temporal dependence in either the unstructured or spatially structured random effects resorting to gaussian random walks. Therefore, information on risks in every region is shared between periods providing more

reliable estimates. This approach has also been used in [20] and [21] in an age-period-cohort frame.

From the classical approach [22] model the spatio-temporal evolution of infant mortality with a CAR distribution to describe the spatial variability in risk, and use a spline for every administrative unit to accommodate its temporal variation. This work incorporates smooth modelling tools to describe the time trend in every location. However, such evolution is considered independent in neighbouring sites, avoiding the transfer of temporal information among these regions, which can be very desirable in the case of rare diseases in order to obtain more reliable estimates.

It is possible to find much more contributions in spatio-temporal modelling in, for example, environmental [23, 24] or ecological applications [25]. Nevertheless, lattice data are not so common in those fields, and geostatistical methods are often used in those areas instead of Gaussian Markov random fields.

As it can be seen, there is neither a spatio-temporal model with the agreement of BYM in the spatial field nor a wide consensus on how to describe the temporal and spatial evolution at a single time in a proper way. In this work a spatio-temporal approach to disease mapping combining time series and spatial modelling ideas is introduced, in order to link risks along time and space at the same time.

This paper is organized as follows: section 2 introduces our autoregressive approach to spatio-temporal disease mapping; section 3 shows some theoretical properties of the approach proposed in section 2 that will provide a useful insight on the dependence structures defined in the former hierarchical model. The results obtained through a simulated dataset and a brief comparison with other models are shown in section 4. Finally, section 5 summarizes the

autoregressive model's behaviour and points out some lines for future development.

2. Autoregressive linking of spatial patterns

The main aim of the approach that we are going to introduce is to define a spatio-temporal structure in which the relative risks are both spatial and temporal-dependent at the same time. On the one side we agree with the philosophy of [12], [14] and [13] of defining a similar temporal evolution in those places that are geographically close. However, our intention is to define a time structure smoother than those with a linear or a quadratic trend, in order to describe any kind of evolution in every geographical unit under study. Therefore, in a way, we also agree with the approach of [22] which may be more appropriate when the number of studied periods is bigger.

Let Y_{ij} be the number of event counts for the ith area and jth period, $i = 1, ..., I; j = 1, ..., J$; E_{ij} the expected number of counts for every area and time interval under the hypothesis that risk remains constant in space and time. This means that, if p_{ijk} is the number of people living in the ith area in the jth season and kth age group, $k = 1, ..., K$; then:

$$E_{ij} = \sum_{k=1}^{K} p_{ijk} \left(\frac{\sum_{i=1,j=1}^{I,J} Y_{ijk}}{\sum_{i=1,j=1}^{I,J} p_{ijk}} \right)$$

where, the kth index for Y_{ijk} refers to the age group.

From now on, we will asume that Y_{ij} follows a Poisson distribution:

$$Y_{ij} \sim Poisson\left(E_{ij} exp(r_{ij})\right) \quad i = 1, ..., I, \ j = 1, ..., J$$

where $exp(r_{..})$ stands for the relative risk for every area and time interval under study. Once we have defined the observed event counts as a function of its expected value and the relative risk, it is necessary to define a structure on r that takes profit of spatial and temporal relations

to get reliable estimates. With that aim we define the log-relative risk at the first observed time period as sum of an intercept, and two random effects, that's to say:

$$r_{i1} = \mu + \alpha_1 + (1-\rho^2)^{-1/2} \cdot (\theta_{i1} + \phi_{i1}),\ i = 1, ..., I \tag{2.1}$$

$$\theta_{i1} \sim \mathcal{N}(0, \sigma_\theta^2),\ i = 1, ..., I$$

$$\boldsymbol{\phi_{1:I,1}} = (\phi_{11}, ..., \phi_{I1}) \sim \text{CAR.normal}(\sigma_\phi^2)$$

where the above distributions are parameterized as a function of their variances. The expression *CAR.normal* stands for an intrinsic Gaussian conditionally autoregressive distribution [11] under the restriction that the sum of all of its components must equal 0, in order to guarantee the property of the resulting distribution on the (n-1)-dimensional resulting space. We will refer to this term as the spatial effect. We include both random effects, spatial and heterogenous, to describe the spatial pattern of risk and to ensure enough flexibility to allow very different risk estimates in close places. In the former expression, ρ corresponds to the temporal correlation that we are going to introduce next. μ models the mean level of risks for all the periods and regions and α_1 models the mean deviation of the risks in the first period from the mean level for all of them. For the following time intervals we define the relative risks as:

$$r_{ij} = \mu + \alpha_j + \rho \cdot (\theta_{i(j-1)} + \phi_{i(j-1)}) + \theta_{ij} + \phi_{ij},\ i = 1, ..., I,\ j = 2, ..., J \tag{2.2}$$

$$\theta_{ij} \sim \mathcal{N}(0, \sigma_\theta^2),\ i = 1, ..., I,\ j = 2, ..., J$$

$$\boldsymbol{\phi_{1:I,j}} \sim \text{CAR.normal}(\sigma_\phi^2),\ j = 2, ..., J$$

$$\boldsymbol{\alpha_{1:I}} \sim CAR.normal(\sigma_\alpha^2)$$

Hence, the expected values for the relative risk in every region and moment will not depend only on its neighbours' estimates during that exact moment, but it will also depend on their

estimates on previous periods. The dependence in every administrative unit among different time periods has been defined as an order 1 autoregressive time series, so that risk estimates are temporally dependent. On the other hand the spatial random effect at every time interval ensures the geographical dependence of risk estimates. Thus, the former modelling enables us to transfer information between neighbouring time periods and regions as we wanted to happen.

The prior distributions used for the parameters defined just above are the following:

$$\sigma_{\phi}^{-2},\ \sigma_{\theta}^{-2},\ \sigma_{\alpha}^{-2} \sim Gamma(a,b)$$

$$\rho \sim \mathcal{U}(-1,1), \quad \mu \sim N(0,c)$$

In them, prior distributions for the mean risk value for all the time periods, and the precisions of the random effects are defined in terms of several hyperparameters intended to express vague information. Regarding with the prior distribution on the temporal correlation parameter, ρ, has been chosen in a way that ensures the stationarity of the time series, considering that it has an order 1 autoregressive structure. For α, the vector that accounts for time evolution in the mean level of risk for every period, a *CAR.normal* proposal has also been used as prior distribution. In this occasion, consecutive periods have been considered as neighbouring, in order to define the dependence structure of the *CAR.normal* distribution. This prior distribution does not rely on any predetermined parametric shape that could condition its values.

In equation (2.1) the term $(1-\rho^2)^{-1/2}$ is introduced in order to make the variance-covariance matrix of $Y_{\cdot 1}$ equal to the stationary covariance matrix of the series $(Y_{\cdot j})_{j=1}^{\infty}$. Thus, the risk distribution for the first period is defined in the same terms than the following ones. Moreover, this selection will enable us to derive several properties of the former spatio-temporal model

in the following section, which will give us some insight about the dependences defined in the model. Nevertheless, it is also possible to use other alternatives to define the variance structure for the first period, as for example, to use precisions for the heterogenous and spatial terms different to those in the latter periods. We think that this is also a very good alternative to the previous modelling as the precision parameters for the first period will focus in the description of the spatial distribution of risks for the first year; while the precision parameter for the subsequent periods will be used to describe the temporal changes on risks that can have very different spatial structure to that of the risks in the first period.

We would like to stress that the temporal increments from one period to the following have spatial structure in the proposed model, so that not only the risk distribution will have spatial structure in our proposal but also the time trend in every administrative unit will also show geographical dependence. Therefore, neighbouring regions will have similar risk evolutions, in the same way that they have similar geographical risk estimates. Thus the estimations of risk evolution will also be based in the information of neighbouring regions, providing more reliable estimations. We will study this fact with more detail in the next section.

Lastly, the log-relative risk evolution on time at every moment and place has been defined as a linear function of such value in the previous season instead of using a trend with a predefined shape for the whole period under study. Therefore, a flexible modelling is used to describe temporal evolutions in a similar way to that used for the geographical term. We hope that this modelling will be able to describe non-linear evolutions of risk in a proper way.

3. Some alternative views to dependence structure

First of all we are going to introduce some of the notation which will be used during this section. In order to reduce the length of the expressions we will denote as $\boldsymbol{\theta_j} = (\theta_{1j}, ..., \theta_{Ij}), j = 1, ..., J$; $\boldsymbol{\phi_j} = (\phi_{1j}, ..., \phi_{Ij}), j = 1, ..., J$ and $\boldsymbol{r_j} = (r_{1j}, ..., r_{Ij}), j = 1, ..., J$. The expression $\boldsymbol{r_j}|\boldsymbol{...}$ for any j will denote the log-relative risk distribution in period j given μ, $\boldsymbol{\alpha}$, σ_ϕ and σ_θ, all of the parameters in the underlying layers of the model. Moreover, $r_{ij}|r_{-(ij)}$ (and analogously $\boldsymbol{r_j}|\boldsymbol{r_{-j}}$) will denote the log-relative risk distribution for the i administrative unit and the j period given the rest of log-relative risks present in the model. Lastly, Σ will denote the variance-covariance matrix of $\boldsymbol{\theta_j} + \boldsymbol{\phi_j}|\sigma_\theta, \sigma_\phi$ for any period j. Note that, for the model stated in the former section, Σ will not change from period to period as its prior structure and hyperparameters remain unchanged along time. Moreover, Σ can also be seen as the sum of a positive definite diagonal matrix of range n, and a positive semidefinite matrix of rank (n-1), which implies that Σ will also be positive definite and thus defines a full rank and invertible variance-covariance matrix.

In the previous section, the log-relative risks for every period have been defined conditioned to those of the previous ones. First of all we will derive the joint distribution of all those risks. In order to get such distribution it will be useful to calculate their covariances, thus:

$$Cov(\boldsymbol{r_j}, \boldsymbol{r_{j+1}}|...) = Cov((\mu + \alpha_j) \cdot \boldsymbol{1_I} + \rho \cdot (\boldsymbol{\theta_{j-1}} + \boldsymbol{\phi_{j-1}}) + (\boldsymbol{\theta_j} + \boldsymbol{\phi_j}),$$

$$(\mu + \alpha_{j+1}) \cdot \boldsymbol{1_I} + \rho \cdot (\boldsymbol{\theta_j} + \boldsymbol{\phi_j}) + (\boldsymbol{\theta_{j+1}} + \boldsymbol{\phi_{j+1}})) = \rho \cdot Cov(\boldsymbol{\theta_j} + \boldsymbol{\phi_j}, \boldsymbol{\theta_j} + \boldsymbol{\phi_j}) = \rho \cdot \Sigma$$

for all the periods j ranging from 1 to $(J - 1)$. Proceeding in a similar way the covariance expressions for all the pairs $\{\boldsymbol{r_j}, \boldsymbol{r_{j'}}\}$ can be calculated for $j, j' = 1, ..., J$; as a consequence

the joint distribution of all the log-relative risks can be derived as:

$$\begin{pmatrix} \boldsymbol{r_1} \\ \boldsymbol{r_2} \\ \vdots \\ \boldsymbol{r_J} \end{pmatrix} |... \sim \mathcal{N}\left(\begin{pmatrix} (\mu+\alpha_1)\cdot \boldsymbol{1_I} \\ (\mu+\alpha_2)\cdot \boldsymbol{1_I} \\ \vdots \\ (\mu+\alpha_J)\cdot \boldsymbol{1_I} \end{pmatrix}, \frac{1}{1-\rho^2} \begin{pmatrix} \Sigma & \rho\Sigma & ... & \rho^{J-1}\Sigma \\ \rho\Sigma & \Sigma & ... & \rho^{J-2}\Sigma \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{J-1}\Sigma & \rho^{J-2}\Sigma & ... & \Sigma \end{pmatrix} \right) \tag{3.1}$$

The former expression can also be expressed in a more summarized form as:

$$\begin{pmatrix} \boldsymbol{r_1} \\ \boldsymbol{r_2} \\ \vdots \\ \boldsymbol{r_J} \end{pmatrix} |... \sim \mathcal{N}\left((\mu \cdot \boldsymbol{1_J} + \boldsymbol{\alpha}) \otimes \boldsymbol{1_I}, \Lambda \otimes \Sigma \right) \tag{3.2}$$

in which $\otimes$ denotes the Kronecker product of two matrices and Λ denotes the correlation matrix of an order 1 autoregressive time series of length J, which means:

$$\Lambda = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & ... & \rho^{J-1} \\ \rho & 1 & \rho & ... & \rho^{J-2} \\ \rho^2 & \rho & 1 & ... & \rho^{J-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{J-1} & \rho^{J-2} & \rho^{J-3} & ... & 1 \end{pmatrix}$$

Thus, we can see how the proposed model structures the temporal and spatial dependence and how the information is shared among the different periods. In particular, the spatio-temporal model just introduced defines a separable covariance structure. Note that the model with independent risks in time is just a particular case of the one stated just above with $\rho = 0$.

In section 3.4.3 of [2] the construction of an Intrinsic Gaussian Random Field as a Kronecker product of matrices is already proposed. Although the present model follows a rather different

formulation to that used in [2], as it is based in the conditional distribution of the risk for every period conditioning to the previous ones, instead of its joint distribution. Nevertheless, expression 3.2 shows that both models have a common covariance structure. In fact, the Type IV interaction defined in [19] as a Kronecker product of two matrices, one defining the spatial structure and another defining the temporal one, can be seen as a particular case of our model without heterogeneous term and with ρ fixed to 1, as he uses a random walk instead of an autoregressive process to link information in time.

Once equation number (3.2) states the joint distribution of the log-relative risks for the whole periods under study, if we condition the distribution in one period to the information in the rest of the periods, it is easy to show that:

$$\boldsymbol{r_j}|\boldsymbol{r_{-j}}, ... \sim \mathcal{N}(\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j}) \tag{3.3}$$

where, if $\boldsymbol{m_j} = (\mu + \alpha_j) \cdot \boldsymbol{1_I}$:

$$\boldsymbol{\mu_j} = \begin{cases} \boldsymbol{m_j} + \rho \cdot (\boldsymbol{r_{j+1}} - \boldsymbol{m_{j+1}}) & j = 1 \\ \boldsymbol{m_j} + \frac{\rho}{(1+\rho^2)} \cdot ((\boldsymbol{r_{j+1}} - \boldsymbol{m_{j+1}}) + (\boldsymbol{r_{j-1}} - \boldsymbol{m_{j-1}})) & j = 2, ..., (J-1) \\ \boldsymbol{m_j} + \rho \cdot (\boldsymbol{r_{j-1}} - \boldsymbol{m_{j-1}}) & j = J \end{cases}$$

and

$$\boldsymbol{\Sigma_j} = \begin{cases} \boldsymbol{\Sigma} & j = 1 \\ (1+\rho^2)^{-1} \cdot \boldsymbol{\Sigma} & j = 2, ..., (J-1) \\ \boldsymbol{\Sigma} & j = J \end{cases}$$

Thus, in our approach, risk estimates for every period have spatial covariance structure, as it is usual in most disease mapping model proposals. Moreover, the mean values for the log-relative risks are not constant in every moment, as they vary geographically depending on the risk estimates for the neighbouring periods. In fact, if a risk excess has been determined

for a region in periods $j-1$ and $j+1$, the expected relative risk for such region in period j will have a value higher than one, and as a consequence a risk excess estimate will be also very likely for this region in period j. This property will give the estimates of consecutive periods a higher coherence, as high/low risk periods in general will be followed by high/low risk periods.

Regarding the distribution of the log-relative risk for the i geographical unit in the j period conditioned to the rest of log-relative risks, it will be useful to rearrange expression (2.2) in the following way:

$$r_{ij} = \mu + \alpha_j + (\rho\theta_{i(j-1)} + \theta_{ij}) + (\rho\phi_{i(j-1)} + \phi_{ij}),\ i = 1, ..., I,\ j = 2, ..., J$$

Thus, risks are expressed as a sum of two autoregressive effects, being the first one heterogeneous and the second one spatially structured.

On the one hand, if we only had the heterogeneous term in the former expression, the log-relative risks for every region would follow an order 1 autoregressive time series geographically independent of its neighbours. On the other hand, in the case of only having the spatial term we would have as:

$$P(r_{ij}|r_{-(ij)}, ...) = P\left((\boldsymbol{r_j}|\boldsymbol{r_{-j}}, ...)_i|(r_{1j}, ..., r_{(i-1)j}, r_{(i+1)j}, ..., r_{Ij}|\boldsymbol{r_{-j}})\right)$$

$P((r_j|\boldsymbol{r_{-j}}, ...)_i)$ is already known, as it has been calculated previously in equation (3.3). And, if we only have the spatial term, which follows a Gaussian Markov Random Field distribution, we have that:

$$P(r_{ij}|r_{-(ij)}, ...) \sim \mathcal{N}\left((\boldsymbol{\mu_j})_i + \frac{1}{n_i}\sum_{k\sim i}(\boldsymbol{r_j} - \boldsymbol{\mu_j})_k, \frac{\sigma_\phi^2}{n_i}\right)$$

where $k \sim i$ denotes all the regions k neighbouring with region i, and n_i denotes its number of neighbours. If the expression of $\boldsymbol{\mu_j}$ is taken into account it can be seen that $r_j - \boldsymbol{\mu_j}$ denotes the deviates of r_j from its predictions based only on its neighbouring periods. Thus, the

expected log-relative risk for r_{ij} depends on the residuals of the temporal predictions of its neighbours. In fact, if the neighbouring regions of i for period j have higher/lower risks than its expected value (attending to its behaviour in their adjacent periods), then r_{ij} will also have an expected log-relative risk higher/lower than zero. Therefore the time trend for every region will also depend on its neighbours' time trends.

In the case of having both spatial and heterogeneous terms, the log-relative risks will follow a combination of the information in such region at different moments, provided by the heterogenous time series, and some information from the time trends from the neighbouring regions. The weight of these two components in the final combination will be determined by the values that σ_ϕ and σ_θ take in their posterior distributions. Moreover, the contribution of the information in the rest of periods to risk estimates at every moment is modulated by the temporal correlation. In that sense, values of $|\rho|$ close to 1 point to a high temporal correlation and much information is shared between consecutive periods; meanwhile, values of $|\rho|$ close to 0 point to temporal independence, avoiding the transfer of information about risks between consecutive periods.

Finally, several properties have been derived for the model stated in section 2. In fact, it has been shown that such proposal defines a separable spatio-temporal covariance structure to describe risk evolution in time and space. Nevertheless, starting from the above model, a wide variety of non-separable covariance structures can be defined without much effort, simply introducing slight modifications in its formulation. One of such modifications has already been mentioned before, this would be to consider the covariance structure for the first period different to those of the following ones. But different precision parameters for all the periods could also be considered and even an autoregressive term could be added for the correlation

parameter that could also vary with time. Thus, it is really straightforward to extend the former model to non-separable spatio-temporal structures that could explain the observed events in a more flexible way.

4. Performance on a simulated dataset

Three simulated datasets have been generated to test the performance of the model introduced in section 2. The 47 mainland provinces in Spain, shown in figure 1, have been considered as the area of study. For everyone of them 15 Poisson counts have been generated intending to represent the observed events in different time periods. The relative risks for those periods are considered to evolve smoothly both in time and space. Indeed, for every time period j, the relative risk in every province i is proportional to the value of the surface:

$$\lambda(x|j) = 1 + \pi_1 \cdot \mathcal{N}_2(x|\mu(j), \Sigma_1) + \pi_2(j) \cdot \mathcal{N}_2(x|Y_2, \Sigma_2) + \pi_3 \cdot \mathcal{N}_2(x|Y_3, \Sigma_3(j))$$

in its centroid. Therefore, the relative risk in every location is proportional to the sum of a constant surface and three additional bivariate normal components, the first one moves as time evolves, the second one changes its contribution to the whole risk as a function of time and the third one changes its dispersion every period. This means that the expectation for the first component moves from location A to B of figure 1 during the 15 simulated periods, the variance-covariance matrix for this component is diagonal with a standard deviation of 100 km both for northing and easting. Finally π_1, which controls the number of events generated by this component, takes a value of 10. The second component keeps still during all the periods in location C of figure 1 with a diagonal and spherical covariance matrix of 100 km of standard deviation. However, its strength changes over time raising from 3 to 10 during the first 8

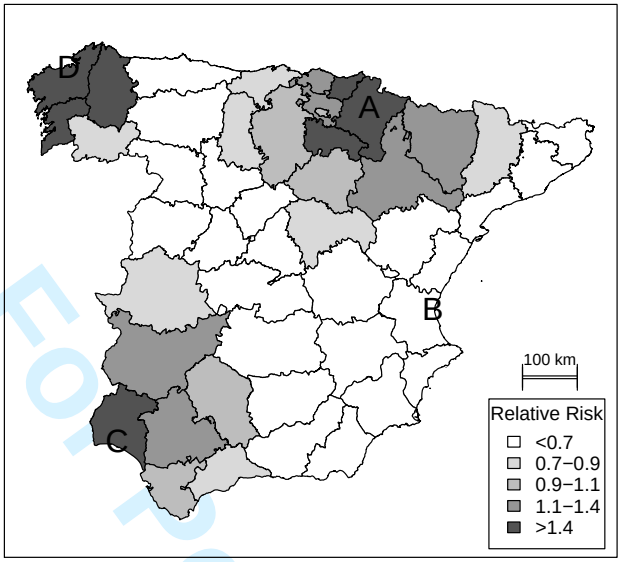

Figure 1. Risks for the mainland provinces of Spain in the first simulated period.

periods and then descends again to 3 during the last periods. Lastly, the third component remains at point D during all the periods in the same way as its strength remains to be 10, however its covariance matrix changes with time. In fact, such matrix is also spherical and diagonal, but its standard deviation increases from 50 kilometers to 200 kilometers from the first to the last period.

The risk for every region and period is proportional to the risk surface estimation in its centroid, more precisely, if X_i $i = 1, ..., 47$ are all those centroids, the number of observed

cases in province i and period j follows a Poisson distribution with mean:

$$\delta \cdot \left(\frac{\lambda(X_i|j)}{\sum_{k=1}^{47} \lambda(X_k|j)/47} \right)$$

Note that the former expression is formed by the product of δ, a parameter that controls the expected number of events for all of the regions, and a second term of mean value equal to 1 which can be interpreted as the relative risk at every region and moment. Three different datasets have been generated following the former mechanism, for δ having values 1, 3 and 5 so that they only differ in the expected number of cases at every region and period but not in their risk distributions. Figure 1 shows the relative risk distribution for the first simulated period of the three datasets. The three former datasets can be found at the web page *http://www.uv.es/mamtnez/ETpaper/datasets.txt*.

For every one of the datasets generated 6 models have been fitted. First, the BYM approach has been considered ignoring the temporal evolution of risk. Secondly, two models with temporally independent effects have been proposed. In them, spatial and independent effects are included for every period, with independent precision terms for every time interval in the first proposal (we will call this model *Ind1*), and sharing the same precision value for the second one (*Ind2*). Finally, three models with autoregressive temporal structure have also been proposed. The first one (*AR1*) corresponds exactly to the model described in section 2. The second one (*AR2*) considers the precision parameter for the first period different to those described in the following periods, thus having a non-separable spatio-temporal correlation structure. The third model (*AR3*) is similar to *AR2* but it only includes the spatial term for every period, leaving out the heterogenous term, in order to get a more parsimonious description of risks behaviour.

Regarding to the hyperparameters of the above models a Gamma distribution with

parameters $a = 0.5$, $b = 0.005$ has been defined for the precisions of the random effects in all the models. On the other hand, a precision of 0.01 has been used for the prior distribution of μ. Inference for all the models has been carried out in WinBUGS 1.4.1 [26] and the syntax for the model described in section 2 can be found at the web page *http://www.uv.es/mamtnez/ETpaper/model.txt*. After discarding 5.000 iterations of burn-in, 10.000 more were generated and only 1 of every 10 of them was saved. Convergence checking was performed attending to the potential scale reduction factor and the effective sample size implemented in the package R2WinBUGS of the statistical software *R* [27]. More information about these statistics can be found in [28]. It has been tested that the potential scale reduction factor has been lower than 1.05 and the effective sample size above 100 for all the parameters in the former models.

Table I shows the correlation and mean squared error (MSE) between the posterior mean of the relative risk and its true value in every region by period combination, for every dataset and model run. The DIC [29] model selection criterium for all the above models is also shown.

Several facts arise from the observation of table I, firstly a great agreement among the 3 criteria used to evaluate the quality of the fit can be seen. In general those models with higher correlations are those with lower MSE and lower DIC, although the last one does not take into account the true relative risk to assess the fit.

On the other hand, the autoregressive models show better fit than the other ones, so that the structuring of temporal correlation seems to have a very desirable effect on fitting terms. The *BYM* proposal, which ignores time trend, shows even better results in our example than those models with temporally independent estimates. Moreover model *Ind2*, which shares the precisions of the random effects among periods, have better performance than model *Ind1*.

Therefore, it seems to be really advisable to share information among different periods, in fact the worst model from those analyzed has been the one in which information was not shared at all. Regarding to the autoregressive temporal modelling, model *AR2* looks slightly better than model *AR1* in those datasets with higher number of observed events, moreover model *AR3* shows better performance than the former two. Those differences are due to the inclusion of the heterogeneous and spatial effects to model the differences from period to period, which can overparameterize the model (unless the amount of observed events is high enough to admit these two sources of variability). This fact can be confirmed in table I as the third model shows better results when the number of events generated is lower (δ=1) than when this number is higher (δ=5).

Table I also suggests that other autoregressive models could be considered. It can be seen that by excluding the heterogenous term and considering the first period as different, improves the fit of the model. Therefore it would be interesting to consider a model with both modifications at a time. Which means the original variability could be described appropriately in the first period with its own precision parameters and both spatial and heterogenous effects, and from that moment the heterogeneous term could be ignored to describe the time trends in a more parsimonious way.

Regarding model performance as a function of the number of observed events, table I shows that those models which share less information among different periods have the worst performance when such number is lower. Indeed, correlation for model *Ind1* decreases from 0.76 to 0.55, for δ equal to 5 and 1 respectively, while such correlation only decreases 0.06 units in the *BYM* and *AR3* models. Thus, the transfer of information among periods is particularly useful when the number of observed events is lower and the information available for every

period is scarcer. On the contrary, the BYM has been tested to be the only model that increases its deviance when the number of observed events increases (result not shown), meaning that this approach is not so good when there is a higher amount of information in every period. In that case a smoother spatio-temporal model can be a better alternative as there is enough information available to carry it out.

From now on, we will focus on the description of the results for the dataset with $\delta = 3$. Table II shows the posterior mean and 95% posterior credibility interval for the temporal correlation parameter. Similar values are obtained for the 3 autoregressive models. Note that model *Ind2* is the particular case of model *AR1* with ρ equal to 0, thus if we observe the posterior mean of ρ in this last model, it looks clear why the data are better fitted by the autoregressive proposal, as it is a generalization of the former model. As the temporal autocorrelation parameter is very high (around 0.9), the risk estimates for every region in consecutive periods will be very similar and a smooth temporal trend will be described by those risks.

On the posterior distribution of the precisions of the random effects in the autoregressive models, we find that in model *AR1* the posterior mean of the spatial precision, σ_{ϕ}^{-2}, is 30.30 and its 95% credibility interval is [11.45,95.48]; in the same way, the heterogeneous precision σ_{θ}^{-2} has a posterior mean of 35.11 [18.64,58.04]. For model *AR2* different precisions for the first period and the following ones are considered, in fact for the first period the posterior estimate for the spatial precision (2.92 [1.14,7.05]) is remarkably lower than for the rest of periods (62.64 [16.16,165.70]). This result is not true for the heterogeneous precision, as the precision for the first period is 41.56 [6.76,174.30] versus 43.39 [21.87,81.56] for the following ones. These values suggest that there are remarkable differences between risk distribution for the first period and the posterior steps in risk from one time interval to the following one,

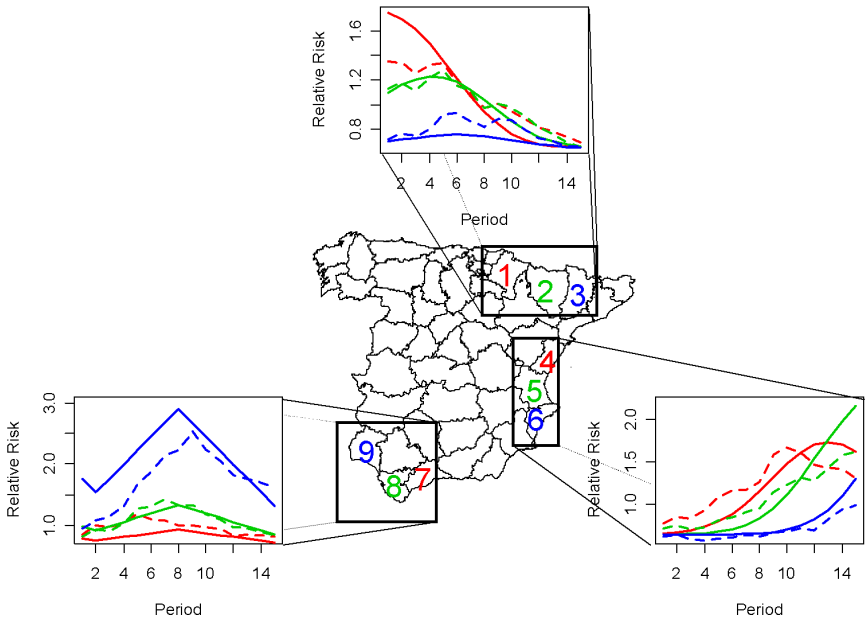

Figure 2. True and estimated time trends for several regions.

as we pointed out in section 2. Finally, the precision for the spatial random effects in model *AR3* is 17.87 [8.95,36.15]. Therefore, if the heterogenous term is removed from the model, the precision for the spatial effect decreases to explain the variability that was previously explained with both components.

Figure 2 displays the risks' time trend for several provinces for model *AR3*, the one with best fit indicators. In the figure, the bold lines stand for the true risk evolution for every selected region and the dashed lines show their predicted trends. As it can be seen, the predicted

trend for every province is a compromise between their true risk values and those predicted in their neighbouring provinces. In general, the predicted trend for every province seems to reproduce the true pattern, although predictions are not too accurate. Nevertheless, more accurate predictions are obtained when the number of observed events is higher (δ=5). It can also be seen that provinces in separate regions have a very distinct behaviour, while those in the same region evolve in a similar manner, which is good from an epidemiological point of view. Time trends also look very smooth due to the presence of temporal correlation, but despite the dependence between consecutive observations, a wide variety of shapes can be adjusted as it can be apreciated in figure 2.

Finally, maps showing the true relative risks and their posterior estimations in models *AR3* and *Ind2* (from the dataset with δ=3) for the 15 periods are displayed at the web page *http://www.uv.es/mamtnez/ETpaper/maps.txt*. These models have been chosen as they are the best in terms of fit between all the independent and autoregressive ones, respectively. We can observe that for these, the *AR3* model has much closer predictions to the true risk than those from the *Ind2*. Moreover, they are temporally much more stable showing a greater coherence from period to period than those from model *Ind2*.

5. Discussion and future lines of development

The model just proposed shows several characteristics that, in our opinion, make it very attractive. We would like to remark the following:

- Information is shared in time in a similar manner to that used to share information in space. Hence, risks are smoothed accounting for both the information in spatial and

temporal neighbours and therefore more reliable estimates are obtained as they are based in a greater amount of information. Moreover, the definition of temporal structure as an autoregressive process enables the data to modulate its importance, in contrast to a Gaussian random walk proposal. Therefore, the balance between the spatial an temporal dependence is determined by the data.

- As it can be seen in the simulated datasets, neighbouring regions have similar time trends. This assumption has been widely used for the geographical risk distribution and its application to the time trend seems very reasonable from an epidemiological point of view.
- Time trends for every region do not rely on a specific parametric shape which has to be specified in the model formulation. Such shape would condition the risk estimates as, in that case, they would have to follow the parametric family imposed in the model and it may not be flexible enough to describe the variety of time trends that arise in the data. This feature of our model is even more important when the number of periods under study is higher and a wider collection of shapes can be described in the geographical units under study.
- The conditional approach just introduced for the temporal structure (instead of its direct definition as a Kronecker product on the joint structure matrix) is in our opinion one of the main advantages of the former model. This new approach permits introducing slight modifications in the model formulation which extend the joint structure matrices of these models to a broader class than that formed by Kronecker product of matrices. Indeed, the joint structure matrix of model *AR2* in this paper is no longer separable, can not be expressed as a Kronecker product of matrices and it seems to have better fit

properties than its analogous to the Kronecker product, model *AR1*.

- This new proposal makes it possible to introduce a very complex covariance prior structure with a reasonable computational burden. Moreover, the conditional approach just introduced enables to make inference in standard bayesian simulation packages, as for example WinBUGS. We think that this is another non negligible practical advantage of this model as it is going to be possible to carry it out without a huge effort by a wide community of potential users.

Regarding to the temporal term, in our simulated dataset it has been shown that to consider such term independently among time for every region, can make the model even worse in terms of fit, mainly when the number of observed events is lower. Nevertheless, sharing information among periods has been shown to improve the fit and makes it possible to describe time trends which in a spatial-only study would be ignored. In our experience with the model in real datasets, we have seen that the temporal dependence is generally very important, indeed higher than the spatial one. Thus, the hypothesis of similar risks in time seems to be at least as reasonable as the widely spread hypothesis of considering similar risks in space.

The proposed model has only a limited predictive power. A parametric temporal approach, as for example the one in [12], would have better predictive properties in general; but, in exchange, the capability to describe any shape in the time trends would be lost and the predictions would rely on the validity of such parametric assumption in the future. The predictive abilities of the proposed model could also be increased by considering a spatio-temporal age-period-cohort model, as for example those in [20] and [21]. Nevertheless, the mere spatio-temporal description of risks during a long period of time has a high interest by its own, in fact it is adding a new dimension to traditional disease mapping to help to determine the risk factors involved in the

studied disease. Moreover spatio-temporal disease mapping methods provide an updated view of the geographical risk distribution instead of the mean risk distribution over a long period of time, thus these methods let to obtain a close risk description at the present time in contrast to habitual disease mapping proposals.

As future lines of development of our work we would like to point out three directions. First, we think that the modifications of the model proposed in section 2 should be explored and it would very interesting to assess the improvements that they provide in terms of fit. For example, it looks interesting to consider different precision parameters for every period or to allow the temporal correlation parameter vary with time. These model modifications generate more flexible covariance structures that could provide a better fit, but as it can be seen in section 4 a balance has to be kept between the flexibility of the model and the amount of information available. In fact, it has been seen that too flexible structures are only acceptable if an acceptable number of events are observed, and we think that this consideration has to be kept in mind when thinking about new models.

As a second line of development it would be very interesting to consider autoregressive temporal models of order higher than 1. The aim of the present work has been to show that autoregressive linking of spatial terms was possible and that such structure could share enough information between consecutive years maintaining enough flexibility to describe temporal evolution without resorting to a parametric shape. Nevertheless, autoregressive models of higher order can be considered, as in some situations they could provide better fit and predictions than the one presented here. In fact, model selection or averaging between autoregressive models of several orders can also be considered.

Lastly, another time series techniques can be considered to link information between different time periods. Thus, a moving average approach looks straightforward to consider, in a similar way to that used in section 2 to define the autoregressive linkage of information. This modelling is more appropriate for diseases that depend on an event that modifies the risk for two (or more in the case of moving average of higher order) periods. It is also possible to consider the time evolution at every site as an integrated time series. Moreover, moving averages, autoregressive and integrated models can be considered at the same time, in such a way that all the ARIMA methodology would be available to link information between consecutive periods. Nevertheless, we would like to point out that although the time series methodology provides several tools to share information in time, the identification of the structure of the time series followed by the data can be a hard task when the number of time periods available is small.

References

[1] A. B. Lawson, *Statistical Methods in Spatial Epidemiology (2nd. edition)*. John Wiley & Sons, 2006.

[2] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory & Applications*. Chapman & Hall/CRC, 2005.

[3] S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical modelling and analysis for spatial data*. Chapman & Hall/CRC, 2003.

[4] P. Elliott, J. C. Wakefiled, N. Best, and D. J. Briggs, eds., *Spatial Epidemiology*. Oxford University Press, 2000.

[5] J. Ferrándiz, J. J. Abellán, A. López, P. Sanmartín, H. Vanaclocha, O. Zurriaga,

M. A. Martínez-Beneito, I. Melchor, and J. Calabuig, "Geographical distribution of the cardiovascular mortality in comunidad valenciana (spain)," in *GIS for Emergency preparedness and health risk reduction* (D. J. Briggs, P. Forer, L. Jarup, and R. Stern, eds.), ch. Geographical distribution of the cardiovascular mortality in Comunidad Valenciana (Spain), pp. 267–282, Springer-Verlag, 2002.

[6] P. Congdon, "A life table approach to small area health need profiling," *Statistical Modelling*, vol. 2, pp. 63–88, 2002.

[7] R. M. Assunção, J. E. Potter, and S. M. Cavenaghi, "A bayesian space varying parameter applied to estimating fertility schedules," *Statistics in Medicine*, vol. 21, pp. 2057–2075, 2002.

[8] J. Ferrándiz, A. López-Quílez, V. Gómez-Rubio, P. Sanmartín, M. A. Martínez-Beneito, I. Melchor, H. Vanaclocha, O. Zurriaga, F. Ballester, J. M. Gil, S. Pérez-Hoyos, and J. J. Abellán, "Statistical relationship between hardness of drinking water and cerebrovascular mortality in valencia: a comparison of spatiotemporal models," *Environmetrics*, vol. 14, no. 5, pp. 491–510, 2003.

[9] S. Banerjee and B. P. Carlin, "Semiparametric spatio-temporal frailty modeling," *Environmetrics*, vol. 14, no. 5, pp. 523–535, 2003.

[10] A. Mollié, "Bayesian mapping of disease," in *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.), pp. 359–380, London: Chapman & Hall-CRC, 1996.

[11] J. Besag, J. York, and A. Mollié, "Bayesian image restoration, with two applications in spatial statistics," vol. 43, pp. 1–21, 1991.

[12] L. Bernardinelli, D. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi, and M. Songini, "Bayesian analysis of space-time variation in disease risk," *Statistics in Medicine*, vol. 14, pp. 2433–2443, 1995.

[13] R. M. Assunção, I. A. Reis, and C. L. Oliveira, "Diffusion and prediction of leishmaniasis in a large metropolitan area in brazil witha bayesian space-time model," *Statistics in Medicine*, vol. 20, pp. 2319–2335, 2001.

[14] D. Sun, R. K. Tsutakawa, H. Kim, and H. Zhuoqiong, "Spatio-temporal interaction with disease mapping," *Statistics in Medicine*, vol. 19, pp. 2015–2035, 2000.

[15] L. A. Waller, B. P. Carlin, H. Xia, and A. E. Gelfand, "Hierarchical spatio-temporal mapping of disease rates," *Journal of the American Statistical Association*, vol. 92, pp. 607–617, 1997.

[16] H. Xia and B. P. Carlin, "Spatio-temporal models with errors in covariates: Mapping ohio lung cancer mortality," *Statistics in Medicine*, vol. 17, pp. 2025–2043, 1998.

[17] A. A. Nobre, A. M. Schmidt, and H. F. Lopes, "Spatio-temporal models for mapping the incidence of malaria in pará," *Environmetrics*, vol. 16, pp. 291–304, 2005.

[18] S. Richardson, J. Abellán, and N. Best, "Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in yorkshire (uk)," *Statistical Methods in Medical Research*, vol. 15, pp. 385–407, 2006.

[19] L. Knorr-Held, "Bayesian modelling of inseparable space-time variation in disease risk," *Statistics in Medicine*, vol. 19, pp. 2555–2567, 2000.

[20] C. Lagazzio, A. Biggeri, and E. Dreassi, "Age-period-cohort models and disease mapping," *Environmetrics*, vol. 14, pp. 475–490, 2003.

[21] V. Schmid and L. Held, "Bayesian extrapolation of space-time trends in cancer registry data," *Biometrics*, vol. 60, pp. 1034–1042, 2004.

[22] Y. C. Macnab and C. B. Dean, "Spatio-temporal modelling of rates for the construction of disease maps," *Statistics in Medicine*, vol. 21, pp. 347–358, February 2002.

[23] S. k. Sahu, A. E. Gelfand, and D. M. Holland, "Spatio-temporal modeling of fine particulate matter," *Journal of Agricultural, Biological and Environmental Statistics*, vol. 11, pp. 61–86, 2006.

[24] A. E. Gelfand, S. Banerjee, and D. Gamerman, "Spatail process modelling for univariate and multivariate dynamic spatial data," *Environmetrics*, vol. 16, pp. 465–479, 2005.

[25] C. K. Wikle, "Hierarchical bayesian models for predicting the spread of ecological processes," *Ecology*, vol. 84, no. 6, pp. 1382–1394, 2003.

[26] D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn, "Winbugs user manual, version 1.4," tech. rep., Cambridge, UK: MRC Biostatistics Unit, 2002.

[27] R Development Core Team, *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[28] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*. Boca Raton: Chapman & Hall/CRC, 2 ed., 2004.

[29] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures

of model complexity and fit (with discussion)," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, pp. 583–641, 2002.

Accompanying Tables

Model	Correlation			MSE			DIC		
	δ=5	δ=3	δ=1	δ=5	δ=3	δ=1	δ=5	δ=3	δ=1
BYM	0.84	0.82	0.78	0.075	0.078	0.104	967.4	910.4	837.5
Ind1	0.76	0.67	0.55	0.104	0.134	0.190	993.7	981.9	886.1
Ind2	0.79	0.74	0.64	0.093	0.112	0.163	967.1	946.7	876.3
AR1	0.91	0.90	0.83	0.041	0.045	0.100	844.3	849.3	825.2
AR2	0.92	0.91	0.82	0.039	0.042	0.111	843.6	847.1	813.3
AR3	0.93	0.93	0.87	0.034	0.035	0.073	829.5	836.3	813.1

Table I. Model comparison in terms of correlation and Mean Squared Error between the original risks and their posterior estimates. The DIC criterium for model selection is also shown.

Model	Posterior mean	95% Posterior C. I.
AR1	0.85	[0.76,0.92]
AR2	0.88	[0.78,0.93]
AR3	0.91	[0.85,0.96]

Table II. Posterior summaries for temporal correlation in the autoregressive models.