A Unifying Modeling Framework for Highly Multivariate Disease Mapping

Botella-Rocamora, P.[1], Martínez-Beneito, M.A..[2,3], Banerjee, S.[4].

[1] Universidad CEU-Cardenal Herrera.

[2] Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunidad Valenciana (FISABIO).

[3] CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain.

[4] Division of Biostatistics - Public Health School, University of Minnesota.

Abstract

Multivariate disease mapping refers to the joint mapping of multiple diseases from regionally aggregated data and continues to be the subject of considerable attention for biostatisticians and spatial epidemiologists. The key issue is to map multiple diseases accounting for any correlations among themselves. Recently, Martinez-Beneito (2013) provided a unifying framework for multivariate disease mapping.

While attractive in that it colligates a variety of existing statistical models for mapping multiple diseases, this and other existing approaches are computationally burdensome and preclude the multivariate analysis of moderate to large numbers of diseases. Here, we propose an alternative reformulation that accrues substantial computational benefits enabling the joint mapping of tens of diseases. Furthermore, the approach subsumes almost all existing classes of multivariate disease mapping models and offers substantial insight into the properties of statistical disease mapping models.

1 Introduction

Public health professionals, including researchers and administrators are often required to map rates pertaining to disease mortality, morbidity, incidence etc. over areal units (e.g. counties, census-tracts) to elicit geographical patterns of disease. When measurements for multiple diseases are recorded at each areal unit, we need to consider *multivariate* areal data models, in order to account for dependence among the multivariate components as well as the spatial dependence between areal units. Multivariate disease mapping continues to be the subject of considerable attention for biostatisticians and spatial epidemiologists and seeks to estimate the risk for a specific disease and location by utilizing information from associated diseases as well as neighbouring areal units. This yields more reliable estimates as compared to traditional univariate disease mapping.

Spatial analysts, therefore, are turning their attention to jointly mapping a large number of diseases to effectively harness the additional information available from diseases associated with each other. While Geographical Information Systems (GIS) and related software can create layers of maps corresponding to each disease separately, technologies to effectively model several diseases jointly, thereby borrowing information across them, are difficult to find. One reason for this is that multivariate spatial disease mapping models become computationally onerous as soon as we have more than two or three diseases. In fact, most of the existing multivariate disease mapping literature restricts itself to two or three diseases. One exception is that in Dobra and Lenkoski [1], who tackle 11 diseases but acknowledge that computations will become unmanageable when the number of geographical units is large. Therefore, computationally feasible multivariate disease mapping models for a large number of diseases and over a large number of geographical units are desirable and can considerably enhance the current utility of mortality atlases.

Gaussian Markov random fields (GMRF) [2] underpin much of the statistical models for disease mapping. Multivariate disease mapping using GMRF's usually proceed from one of the following two premises. The first is factor modelling [3, 4, 5, 6], which assumes the existence of a set of underlying factors determining the spatial distributions of the diseases. This approach includes, as special cases, other models proposed in the literature such as the shared component models [7] or SANOVA [8]. The second premise specifies multivariate spatial distributions [9, 10, 11, 12, 13]. Here, a valid multivariate joint distribution is defined for the multiple diseases. This multivariate distribution considers both spatial dependence within diseases as well as association between diseases. Care is taken so that the a legitimate (symmetric and positive definite) covariance matrix is derived to model such dependences.

Recently, Martinez-Beneito [14] offered a general framework for multivariate disease mapping that encompasses a diverse range of statistical models for mapping multiple diseases. In particular, this general approach subsumes either of the two approaches mentioned above. The framework is rich, includes both separable as well as non-separable covariance structures, and can accommodate different spatial dependence structures with different covariance matrices within diseases. However, the approach in Martinez-Beneito [14] is computationally demanding and the number of floating point operations (flops) rapidly increase with the number of diseases in the analysis. Our current work seeks to extend the above work to multivariate disease mapping settings where we encounter a large number of diseases. We do so by developing a simpler and computationally more convenient form that can handle a considerably large number of diseases even within conventional Bayesian simulation packages such as `WinBUGS` [15].

This article is organized as follows. Section 2 briefly discusses the framework in Martinez-Beneito [14] and proposes a more convenient alternative reformulation. Section 3 demonstrates the proposed methodology with some real examples. Here, we test the numerical performance and compare several models implemented by the proposed approach. We then describe the results of our proposal when applied to the study of a large set of causes of mortality. Finally, Section 4 concludes the paper with some further discussion of the results.

2 A computationally convenient proposal for multivariate disease mapping

A general statistical framework for the multivariate disease mapping problem can be described as follows. Let O_{ij} and E_{ij} denote, respectively, the number of observed and expected cases for the i-th geographical unit of study and j-th health outcome, which, from now on, will refer to a specific disease. The data likelihood assumes that

$$O_{ij} \sim Poisson(E_{ij}RR_{ij}) \;\; i = 1, ..., I, \;\; j = 1, ..., J$$

where RR_{ij} is the relative risk for the j-th disease in the i-th geographical unit and is modeled as $\log(RR_{ij}) = \mu_{ij} + \theta_{ij}$. The term $\mu_{ij} = \mathbf{X}_{i\cdot}\boldsymbol{\beta}_j$ introduces covariates that could be having an effect on the log-relative risks. On the other hand, θ_{ij}'s are a collection of random effects whose joint distribution specifies how dependence within and between-diseases is defined. For the remainder of this article, we will consider this framework for multivariate disease mapping, focusing on the modeling of the θ_{ij}'s.

2.1 The QR-based multivariate disease mapping model

Let $\boldsymbol{\Phi}$ be an $I \times J$ matrix of zero-mean Gaussian random effects, whose j-th column $\boldsymbol{\phi}_j$ has a marginal spatial covariance matrix $\boldsymbol{\Sigma}_j$. Collecting the θ_{ij}'s into an $I \times J$ matrix $\boldsymbol{\Theta}$, Martinez-Beneito [14] assume that

$$\boldsymbol{\Theta} = \boldsymbol{\Phi}\left(\widetilde{\boldsymbol{\Sigma}}_b\mathbf{P}\right)^\top = \boldsymbol{\Phi}\mathbf{P}^\top\widetilde{\boldsymbol{\Sigma}}_b^\top \tag{1}$$

where $\mathbf{P}$ is some orthogonal matrix and $\widetilde{\boldsymbol{\Sigma}}_b$ is the lower-triangular factor in the Cholesky decomposition of the between-diseases covariance matrix $\boldsymbol{\Sigma}_b$. The structure in (1), which we refer to as the **QR**-model, is highly general as it includes almost any GMRF-based multivariate disease mapping models in current use. For example, suppose that our map comprises I regions and consider a simple $I \times I$ binary adjacency matrix $\mathbf{W}$ with zeroes along the diagonal and whose (i, i')-th off-diagonal entry is 1 if i and i' are neighbors, and 0 if they are not neighbors. Let $\mathbf{D}$ be an $I \times I$ diagonal matrix, whose diagonal entries contain the the number of neighbors for each region. If each column of $\mathbf{\Phi}$ has a proper CAR distribution, i.e., $\phi_j \overset{ind}{\sim} N(\mathbf{0}, (\mathbf{D} - \rho\mathbf{W})^{-1})$, where ρ is a smoothness parameter constrained conveniently in $(0, 1)$ to ensure positive definiteness of the covariance matrix, then (1) can be shown to be equivalent to Case 3 of the so called "coregionalized MCAR" model in Jin, Banerjee et al. [12]. Furthermore, unlike the coregionalized MCAR models, (1) may also be implemented in the `BUGS` language [14] making it accessibile to a much wider community of users.

In spite of its generality, (1) reveals some potentially unattractive features. First, the orthogonal matrix $\mathbf{P}$ is constructed by a composition of Givens rotation matrices [16]. Each of these rotation matrices are parametrized by a set of angles, say $\{\psi_{jj'}, 1 \leq j < j' \leq J\}$, and effects a rotation by $\psi_{jj'}$ about one of the main axes in the J-dimensional space defined by the set of diseases in the study. A Uniform prior distribution is usually assumed for each $\psi_{jj'}$ but this choice does not deliver a Uniform prior distribution on the set of orthogonal matrices, which would seem to be a natural choice for $\mathbf{P}$. Moreover, the sensitivity of this specification to the prior distributions on the Givens angles is unclear.

Second, and very pertinent to our current application, the $\mathbf{QR}$-based model suffers from scalability issues that preclude its implementation for multivariate studies with a modestly large number of diseases (say more than 5, although it will also depend on the number of areal units in the application). To be precise, a valuable feature of (1) is that, by construction, it ensures legitimate probability models but it entails sampling from a Uniform distribution on the cone of symmetric, positive-definite matrices. Martinez-Beneito [14] achieved this by sampling correlations among diseases from a Uniform distribution between -1 and 1 and subsequently checking for positive-definiteness of the corresponding covariance matrix. The set of positive-definite matrices resides in a much smaller part of the $[-1, 1]^{J(J-1)/2}$ cube when J, the number of diseases considered, is higher [17]. Therefore, the sampling of these matrices becomes inefficient for the study of a large number of diseases. Moreover, in a Bayesian setting, a total of $J(J-1)/2$ Givens angles are drawn at every step of the Markov chain Monte Carlo (MCMC) algorithm in order to sample from the random matrix $\mathbf{P}$. Thus, this quantity grows at a quadratic rate with the number of diseases. Finally, the construction of $\mathbf{P}$ from the Givens angles also becomes much more complex when the number of diseases is greater than three or four. Consequently, some improvements should be made to (1) and its original implementation in order to make it feasible for studying a large number of diseases.

2.2 The M-based modelling proposal

While the $\mathbf{QR}$ decomposition is widely used in statistical modeling and computation, we provide a brief review in the context of the reformulation of (1) that we seek. Any $n \times m$ real matrix $\mathbf{M}$ with $n \geq m$ can be decomposed as the product $\mathbf{M} = \mathbf{QR}$, where $\mathbf{Q}$ is $n \times m$ with orthogonal columns and $\mathbf{R}$ is $m \times m$ upper triangular. If $\mathbf{M}$ has full column rank and we require $\mathbf{R}$ to have positive diagonal entries, this $\mathbf{QR}$ decomposition of $\mathbf{M}$ is unique [18]. Since $\mathbf{M}^\top\mathbf{M} = (\mathbf{QR})^\top(\mathbf{QR}) = \mathbf{R}^\top\mathbf{Q}^\top\mathbf{QR} = \mathbf{R}^\top\mathbf{R}$, we can deduce that $\mathbf{R}$ coincides with the upper-triangular factor of the Cholesky decomposition of $\mathbf{M}^\top\mathbf{M}$.

The $\mathbf{QR}$ decomposition is relevant because the last two terms in the right hand side of (1) have exactly the same form as the $\mathbf{QR}$ decomposition of a general matrix $\mathbf{M}$, namely, the product of an orthogonal and an upper triangular matrix. Hence, we refer to (1) as the $\mathbf{QR}$-based model and we rewrite (1) as

$$\mathbf{\Theta} = \mathbf{\Phi}\mathbf{M}\ , \tag{2}$$

where $\mathbf{M}$ is a nonsingular $J \times J$ matrix. We refer to (2) as the ***M**-based model* where $\mathbf{M}$ will be understood to be a nonsingular but otherwise arbitrary matrix.

The $\mathbf{QR}$-based and $\mathbf{M}$-based models are equivalent except for differences arising in their respective prior specifications. In (1), we require a prior distribution for the orthogonal matrix $\mathbf{P}$ and $\mathbf{\Sigma}_b$, while in (2) a prior distribution for $\mathbf{M}$ is sought. These specifications have repercussions in computing and implementation. The former compels the user to undertake specific inference for both components of the $\mathbf{QR}$ decomposition, i.e. $\mathbf{P}$ and $\widetilde{\mathbf{\Sigma}}_b^\top$ in every iteration of an MCMC algorithm, while the latter enables inference on the single entity $\mathbf{M}$. Thus, the latter strategy does not require the repeated computation of either the Cholesky decomposition of the between-diseases covariance matrix or the corresponding orthogonal transformation matrix. This has been replaced by finding a suitable class of priors on $\mathbf{M}$, to which we return shortly.

The j-th column of $\boldsymbol{\Theta}$, say $\boldsymbol{\theta}_j$, contains the spatially referenced log-risks for the j-th disease. One can, therefore, interpret (2) as positing the vector of log-risks to be a linear combination of underlying latent variables whose coefficients form the j-th column of $\mathbf{M}$. Thus,

$$\boldsymbol{\theta}_j = \boldsymbol{\phi}_1 m_{1j} + \cdots + \boldsymbol{\phi}_I m_{Ij} \ , \tag{3}$$

where m_{ij} is the (i, j)-th entry in $\mathbf{M}$. Expression (3) suggests that the original $\mathbf{QR}$-model can also be viewed as a factor model with non-orthogonal loadings (the columns of the non-orthogonal matrix $\mathbf{M}$) on the different dimensions extracted in the analysis. As a consequence, we can achieve the same degree of complexity in the modelling following either a factor modelling approach or a multivariate-distribution-based approach. Put differently, as just shown, we can regard these seemingly different approaches as two faces of the same coin.

Expression (3) can also accommodate settings where the number of underlying patterns is different from the number of diseases. Clearly, the number of patterns could be chosen to be lower than the number of diseases being modeled. This model can possibly be more restrictive but will benefit from the parsimony (and perhaps provide comparable or better fit) if indeed a lower number of factors could fully explain the diseases in the study. On the other hand, we could also consider a number of underlying factors higher than the number of diseases. This setting corresponds to the convolution process of Besag, York and Mollié [19] in univariate disease mapping problems. In our multivariate setting, the use of more underlying factors than diseases could improve the fit when the family of spatial distributions considered for the columns of $\boldsymbol{\Phi}$ was not appropriate to describe the spatial pattern of one (or some) of the diseases studied. In that case, several columns of $\boldsymbol{\Phi}$, corresponding to spatial patterns of different features (such as proper CAR distributions of different correlation parameters), could be used to model the spatial distribution of the disease(s) providing a more flexible class of spatial distributions (a mixture model) than the original prescription for modeling the columns of $\boldsymbol{\Phi}$.

2.3 Some theoretical properties of the M-based model

As mentioned earlier, the **QR**-based and **M**-based models are equivalent, so their model fitting properties are essentially the same. They are guaranteed to be well defined [14] and they lead to the same covariance structure but with alternate parametrizations. Nevertheless, a few remarks on the covariance structure are warranted. If different covariance matrices, say $\{\mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_J\}$ were used to define the spatial structure for the columns of $\mathbf{\Phi}$, then the covariance matrix of vec($\mathbf{\Theta}$) yields

$$\begin{aligned}
&(\mathbf{M}^\top \otimes \mathbf{I}_I)\text{diag}(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_J)(\mathbf{M} \otimes \mathbf{I}_I) \\
&\quad = (\mathbf{R}^\top \otimes \mathbf{I}_I)(\mathbf{Q}^\top \otimes \mathbf{I}_I)\text{diag}(\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_J)(\mathbf{Q} \otimes \mathbf{I}_I)(\mathbf{R} \otimes \mathbf{I}_I) \\
&\quad = (\mathbf{R}^\top \otimes \mathbf{I}_I)\text{Block}\left(\left(\sum_{k=1}^{J} q_{ik}q_{jk}\mathbf{\Sigma}_k\right)_{i,j=1}^{J}\right)(\mathbf{R} \otimes \mathbf{I}_I)\ , \qquad (4)
\end{aligned}$$

where **Q** and **R** denote the elements of the **QR** decomposition of **M** and q_{ij} denotes the (i, j)-th element of **Q**. Note that in terms of the **QR**-based model $\mathbf{Q} = \mathbf{P}^\top$ and $\mathbf{R} = \widetilde{\mathbf{\Sigma}}_b^\top$. The resulting covariance structure coincides with the most general model in Martinez-Beneito [14] and, if each of the columns of $\mathbf{\Phi}$ follow independent proper CAR distributions with different parameters, with the Case 3 coregionalized MCAR model in Jin, Banerjee et al. [12].

Expression (4) reveals the effect of each of the constituent matrices on the general covariance matrix. The matrix **Q** transforms the J original columns in $\mathbf{\Phi}$ into J linear combinations of those columns. Each column of the resulting matrix has a covariance matrix with the same structure: $\sum_{k=1}^{J} q_{ik}^2\mathbf{\Sigma}_k$, which is a convex linear combination of the covariance matrices of the columns of $\mathbf{\Phi}$. On the other hand, the off-diagonal blocks in that matrix, $\sum_{k=1}^{J} q_{ik}q_{jk}\mathbf{\Sigma}_k$ (the spatial covariances) are just a linear combination of the original covariance matrices whose coefficients add up to 0. For $\mathbf{\Sigma}_1 = \cdots = \mathbf{\Sigma}_J$ the orthogonality of **Q** makes those blocks exactly equal to 0. Instead, if these matrices were different, all of the $\mathbf{\Sigma}_j$'s would reproduce a spatial structure based on the same geographical distribution, so these blocks would not be expected to be very different to 0. Therefore, **Q** broadens the spatial structure of the columns in $\mathbf{\Phi}$ defining a set of weakly dependent patterns of richer spatial structures than the original columns in $\mathbf{\Phi}$. Finally, the $\mathbf{R}$ matrix is in charge of inducing dependence between the weakly dependent enhanced spatial patterns that we have just built using $\mathbf{Q}$. The combined effect of $\mathbf{Q}$ and $\mathbf{R}$ is executed through $\mathbf{M}$.

We now turn to prior distributions for $\mathbf{M}$. As already mentioned both the $\mathbf{QR}$-based and $\mathbf{M}$-based models define identical multivariate spatial distributions. Therefore, the only possible source of discrepancy between them may be the prior specifications for $\mathbf{M}$ in the $\mathbf{M}$-based model and for $(\mathbf{P}, \mathbf{\Sigma}_b)$ in the $\mathbf{QR}$-based model. Can we, then, establish a correspondence between a prior distribution for $\mathbf{M}$ and some equivalent prior distributions for $(\mathbf{P}, \mathbf{\Sigma}_b)$? We now explore this.

As alluded to earlier, the entries in $\mathbf{M}$ can be regarded as coefficients in the regression of the log-relative risks on the (unknown) underlying patterns captured in $\mathbf{\Phi}$. Therefore, it seems reasonable to put independent vague prior distributions on them, such as improper Uniform or zero-centered Normal priors with a large variance. Let us assign $N(0, \sigma^2)$ priors to the entries in $\mathbf{M}$ and consider the improper Uniform prior as a limiting case (σ^2 tending to infinity) of this setting. Then,

$$\mathbf{\Sigma}_b = \mathbf{R}^\top \mathbf{R} = \mathbf{R}^\top \mathbf{Q}^\top \mathbf{Q} \mathbf{R} = \mathbf{M}^\top \mathbf{M} \sim Wishart(J, \sigma^2 \mathbf{I}_J) \ .$$

Moreover, since the elements of $\mathbf{M}$ are i.i.d. zero-centered Normal random variables, $\mathbf{Q}$ will follow a Haar distribution, i.e. a Uniform distribution on the set of orthogonal matrices (page 169 in Gentle [16]). Therefore, an $\mathbf{M}$-based model with zero-centered vague Normal prior distributions for every entry is equivalent to the $\mathbf{QR}$-based model with a Uniform (Haar) prior distribution on $\mathbf{Q}$ and a Wishart$(J, \sigma^2 \mathbf{I}_J)$ distribution for the covariance between diseases.

An alternative specification treats the entries in $\mathbf{M}$ as independent Gaussian random effects. Now σ is assigned a vague Uniform prior between 0 and a large number. This is especially attractive for studying a large set of mortality causes so that $\mathbf{M}$ has a large number of entries (typically hundreds). Modeling them as random effects allows borrowing of strength (shrinkage) from a common distribution and delivers improved inference.

Hence, we explore two different $\mathbf{M}$-based models – one that treats the elements in $\mathbf{M}$ as fixed effects and another that treats them as random effects.

Finally, we remark that the $\mathbf{M}$-based model enables inference for some parameters in the model but not for others. For example, while the log-relative risks in the model ($\mathbf{\Theta}$) or the covariance matrix between diseases ($\mathbf{\Sigma}_b = \mathbf{M}^\top\mathbf{M}$) can be identified, inference for $\mathbf{\Phi}$ or $\mathbf{M}$ is precluded because for any orthogonal matrix $\mathbf{U}$ we have

$$\mathbf{\Theta} = \mathbf{\Phi}\mathbf{M} = \mathbf{\Phi}\mathbf{I}\mathbf{M} = (\mathbf{\Phi}\mathbf{U})(\mathbf{U}^\top\mathbf{M}) = \mathbf{\Phi}^*\mathbf{M}^* \ ,$$

where $\mathbf{\Phi}^* = \mathbf{\Phi}\mathbf{U}$ is composed of a set of spatial effects and $\mathbf{M}^* = \mathbf{U}^\top\mathbf{M}$ is also a general $J \times J$ matrix. Hence, every orthogonal transformation of the columns of $\mathbf{\Phi}$ and the equivalent transformation of the rows of $\mathbf{M}$ yields an alternative decomposition of the log-relative risks $\mathbf{\Theta}$, implying lack of identifiability. Therefore, we avoid specific inference on $\mathbf{\Phi}$ and $\mathbf{M}$ and treat them as data augmentation tools to induce dependence on the log-relative risks. This would be similar to the label switching problem in Bayesian mixture models [20], where the posterior distribution is multimodal with one mode for any permutation of the components of the mixture. This estimation problem does not preclude the estimation of the mixture model because the distribution arising from the mixture is perfectly identifiable. Nevertheless, the identification of their components requires much more care and may require the use of some specialized techniques [20]. Our case would be similar, since the separate estimation of $\mathbf{\Phi}$ and $\mathbf{M}$ would be problematic but inference on $\mathbf{\Phi}\mathbf{M}$ is legitimate because its product is identifiable. Therefore, if no specific MCMC post-processing procedure is used to identify both $\mathbf{\Phi}$ and $\mathbf{M}$, specific inference on these two matrices should be avoided.

3 Analysis of Comunitat Valenciana's mortality data

We now implement and assess the $\mathbf{M}$-based model in studying the geographical distribution of mortality in Comunitat Valenciana, one of the seventeen regions that constitute Spain, with around 5.1 million inhabitants in 2012. The administrative unit used for this study is the municipality, a division ranging from 21 to about 750.000 inhabitants per unit. A total of 540 municipalities make up Comunitat Valenciana. The mortality data used here correspond to the Spatio Temporal Mortality Atlas of Comunitat Valenciana [21] comprising the period 1987-2006. In the said atlas, a total of 46 different causes of mortality were independently studied (23 for men and 23 for women). We illustrate the benefits of multivariate modelling for this dataset.

An added attraction of our **M**-based model is its easy MCMC implementation in the `BUGS` language. All the models we consider here were run on `WinBUGS` [15] and the available code is available from `http://www.uv.es/~mamtnez/Mmodel.html`. For each model, we ran three chains up to $30,000$ iterations per chain. Of these, the first $5,000$ were discarded as burn-in and the resulting chains were thinned to retain one of every 75 iterations due to the large amount of variables to be saved (11,340 in our largest setting). Therefore, $1,002$ (334×3) iterations were saved. The chains used for every model were run in parallel to expedite computations. Instead of computing all three chains within a single call to `WinBUGS`, we made three different parallel calls (one for each chain) using an `R` [22] function developed for this purpose. Thus, each chain was run in a different core of the processor(s) instead of running all three in a single core (as is the default `WinBUGS` call), speeding up simulations in a factor close to the number of chains run. Convergence was assessed by means of visual inspection of the history of the simulated chains, the Brooks-Gelman-Rubin statistic and the effective sample size. Convergence was checked for the Deviance, the vector $\boldsymbol{\mu}$ and the elements of matrix $\boldsymbol{\Theta}$. The simulated chains yielded virtually independent posterior draws with first-order autocorrelations very close to 0 and effective sample sizes close to 1000, the number of iterations saved.

In the next subsection we analyze the Comunitat Valenciana's mortality data using the **M**-based model and assess its performance. We explore different sets of causes of mortality in order to illustrate the modelling possibilities and the performance of several models within our framework. Subsequently, in the second subsection, we illustrate the performance of the model on the analysis of a relatively large dataset that considers the joint study of 21 causes of mortality altogether for the whole Comunitat Valenciana (21 causes · 540 municipalities=11,340 observations). We also explore the feasibility of the random/fixed effects modeling of matrix **M** on datasets of this size.

3.1 Performance of the M-based model

We carried out an extensive analysis on Comunitat Valenciana's mortality data in order to explore some features of the **M**-based model. First, we have compared the computing times for several **QR** and **M**-based models. We have carried out several studies with different numbers of diseases for these two approaches. The diseases analyzed are not listed since these are less relevant for this part of the study. The spatial structure used for all these models was a proper CAR distribution with different spatial correlation parameters.

[PUT TABLE 1 HERE]

We have run the **QR**-based model only for the study of 2 and 3 diseases since some practical problems were found when implementing that model for more diseases. Even for the model with just four causes of mortality, `WinBUGS` rejected the expression of the **Q** orthogonal matrix as 'too complex' to be run. Thus, for 4 or more diseases, it would be necessary to use an alternative software or an alternative implementation to carry out the inference. Nevertheless, even for 2 and 3 diseases, Table 1 clearly reveals scalability problems for the **QR**-based model in comparison to the **M**-based alternative. For the **QR**-based model, analyzing 2 or 3 diseases increases computing times by 162% while for the **M**-based model that increase is 'just' 97%. Computing times for the **M**-based model seem to follow a quadratic trend. For these models, the number of elements to be estimated in the **M** matrix is also a quadratic function of the number of diseases. Hence, the sampling of these values could be consuming a substantial part of the computing time for the MCMC algorithm. In any case, computing times seem affordable when the number of diseases are in the twenties, thereby eliciting the importance and practical benefits of the **M**-based model in problems of such dimensions.

Besides, we have carried out a second analysis on Comunitat Valenciana's mortality data. We have randomly chosen 7 different causes of mortality for men. The studied causes are listed as {Oral Cancer, Larynx Cancer, Cirrhosis, Bladder Cancer, Stomach Cancer, Colon Cancer, Rectum Cancer}, although this selection should not have any large effect on the final results. We have implemented several **M**-based models for these causes of death in order to illustrate its modelling possibilities and to gain some insight on the effect of those modifications. As already mentioned the **M**-based proposal can also be seen as latent factor modelling and, therefore, different numbers of factors could be considered. These models would arise from the modification of the number of rows in the **M** matrix and the number of columns in $\mathbf{\Phi}$. For our analysis, we have explored model performance of the **M**-based model for factors varying between 2 and 9. We have also implemented a BYM spatial structure [19] of 7 underlying factors as an alternative spatial structure to that used in the said proper CAR factor modelling. This model assumes the existence of 14 underlying patterns, 7 spatially heterogeneous and 7 spatially structured with Intrinsic CAR distributions. In this case, the corresponding **M** matrix mixes these 14 patterns to induce dependence between diseases. Finally, a set of 7 independent BYM models has also been implemented in order to assess the improvement achieved by following the multivariate approach. Table 2 shows the Deviance Information Criterion (DIC) model selection criteria [23] corresponding to every one of these models.

[PUT TABLE 2 HERE]

Table 2 shows some interesting results. Factor models with a low number of underlying spatial effects are clearly poorer than those with a number of factors close to the number of diseases being studied. This suggests avoiding the use of low-dimensional factor models for multivariate modelling, even for datasets with a moderate number of diseases. On the other hand, the **M**-based model appears to be a convenient option as it enables the fitting of spatial factor models with several factors, when these models have a lot of practical problems to be fitted. Table 2 also suggests the use of as many proper CAR underlying factors as diseases. Nevertheless, the BYM spatial structure (combining 14 underlying patterns) shows a better fit than the proper CAR model with 7 factors. This suggests that the combination of several underlying factors, corresponding to different distribution families, may improve the fit in general terms. As a consequence, the use of wider classes of spatial processes yielding more general and flexible spatial distributions [24] seems a promising way to improve the fit also in multivariate models.

Finally, the DIC of the disease-independent model is about 150 units higher than the multivariate BYM model, which shows considerable improvement from sharing information between diseases. In this regard, Figure 1 shows the geographical patterns fitted for the BYM independent modelling (upper row) and the BYM multivariate **M**-based model (lower row) for Cirrhosis, Bladder Cancer and Rectum Cancer, respectively. Correlations between the effects for these diseases were 0.94, 0.87 and 0.70, respectively, the maximum, median and minimum correlations attained for all 7 diseases in the study. These patterns from the multivariate study are less smoothed and depict geographical risk variations in a more precise way than those from the univariate models. For some diseases, such as Rectum Cancer, the improvement attained from multivariate modelling is particularly pronounced.

[PUT FIGURE 1 HERE]

3.2 Multivariate study of mortality in men

We turn to the feasibility of the **M**-based model for high-dimensional multivariate disease mapping, the original motivation of this work. Multivariate disease mapping with more than four or five diseases modelled jointly have rarely been explored due to the obvious computational bottlenecks. Here we consider multivariate modelling of mortality along Comunitat Valenciana, a large region comprising 540 municipalitites, for 21 different causes of mortality. These correspond to all the male mortality causes studied in the Spatio-temporal Mortality Atlas of Comunitat Valenciana [21], except for All Cancers and Colorectal Cancer, which are simple combinations of other causes in the study. We also carried out an alternative study of the 12 tumoural mortality causes of the 21 previous causes. This second study was intended to be an example including a more homogeneous set of diseases than the 21 original ones, where the multivariate analysis could make even more sense. Computing times for these models were already included in Table 1 to illustrate the scalability of the $\mathbf{M}$-based model.

Table 3 shows DICs for both studies and for both fixed and random effects modeling of $\mathbf{M}$, for comparative purposes. For the cancer-specific mortality analysis (12 diseases), the random effects model clearly outperforms the one with fixed effects in terms of DIC. However, for the general mortality analysis (21 diseases) one model does not clearly outperform the other. These results point out that if the diseases being studied are similar in some sense, it may be worth modelling the elements in matrix $\mathbf{M}$ as random effects. That analysis could exploit the similarity among diseases and thus improve the fit of the models. On the other hand, if the diseases in the study are more heterogeneous, the benefit of considering the elements of $\mathbf{M}$ as random effects is not so evident. Nevertheless, the main estimates ($\mathbf{\Theta}$ and $\mathbf{\Sigma}_b$) derived from both models are similar and no important discrepancy can be detected on them, from a visual inspection (results not shown).

[PUT TABLE 3 HERE]

Figure 2 shows, in its lower-triangular side, the posterior mean of the correlations between diseases for all types of cancer in the 12-disease study (right and left oriented ellipses hold for positive and negative correlations respectively), derived from the posterior draws of $\mathbf{\Sigma}_b = \mathbf{M}^\top\mathbf{M}$. The upper-triangular portion shows the 80% posterior credible intervals for all these correlations. We find a large proportion of significant correlations whose credible intervals do not include 0. For all those pairs of variables, a common underlying risk factor may exist that determines the geographical pattern of both diseases. All the correlations show posterior means higher than 0, but with very different values. Thus, if one of the municipalities of Comunitat Valenciana has a high mortality risk for one of the tumours studied, it will tend to also have high risks for the remainder of tumours. We can also notice that those pairs of diseases with a higher correlation in Figure 2 are: Lung/Bladder, Lung/Oral, Larynx/Oral, Bladder/Oral, respectively, with correlations higher than 0.6. Most of the deaths for these diseases may be attributed to tobacco consumption; therefore, these results seem to be very reasonable.

[PUT FIGURE 2 HERE]

Results for the 21-diseases study are not shown due to space limitations, although they are reported as supplementary material to the paper. In the 21-diseases version of Figure 2 it can be appreciated that correlations between cancers for that study are similar to those in the 12-diseases study, standing out the same pairs of diseases for both of them (Lung/Bladder, Lung/Oral, Larynx/Oral, Bladder/Oral and so on). Moreover, COPD and Pneumonia, also show high correlations in general with all causes of mortality, mainly with cancer causes. This fact reinforces the idea that the common geographical pattern previously evidenced may be heavily determined by tobacco consumption. Finally Atherosclerosis and Other Cardiovascular Causes are causes showing lower correlations with the rest, which may not be surprising since these two diseases stand out as aetiologically different from the others in the study.

4 Conclusions

This paper proposes an efficient reparameterization of the multivariate modelling proposal in Martinez-Beneito [14]. The modelling proposed can be implemented in `WinBUGS` yielding an efficient implementation and a much simpler coding within this (and other) software(s), avoiding the use of Givens angles or a specific modelling of orthogonal matrices. Nevertheless, the benefits of this formulation go beyond these practical issues. The $\mathbf{M}$-based modelling with a Gaussian $\mathbf{M}$ matrix has been found to be equivalent to the $\mathbf{QR}$ based proposal with a Uniform distribution on $\mathbf{Q}$ and a Wishart distribution for the covariance matrix between diseases. Therefore, the prior distribution of $\mathbf{Q}$, implicit in this approach, seems to be much more reasonable than that arising from a composition of Givens angles, which is difficult to model or to assess its influence on the final results.

Moreover, the modeling framework is versatile enough to subsume a number of specific multivariate disease mapping models that exist in the literature, as discussed in Section 2.

The primary contribution of this paper, we feel, is that it enables the analysis of multivariate geographical datasets with a moderately large set of causes of mortality (tens of causes altogether). Such multivariate analysis will accommodate two separate goals. First, it will allow us to study diseases of low mortality for which univariate studies would yield a poor depiction of the geographical distribution of the risk of that disease. In these cases the multivariate dependence with a large set of supplementary diseases will make the estimated spatial risk surfaces much more meaningful and rich. Second, multivariate studies will allow us to determine relationships between diseases highlighting the existence of common risk factors or unknown associations between some diseases. This research is anticipated to make these goals possible for larger datasets, leading to richer studies and richer epidemiological conclusions. We expect this proposal will enable a new kind of mortality studies/atlases jointly considering a large set of causes of mortality as a whole, instead of as unconnected patterns without any relationship.

Acknowledgments

Paloma Botella-Rocamora acknowledges the financial support of Ministerio de Educación, Cultura y Deporte, via the Programa Nacional de Movilidad de Recursos Humanos del Plan Nacional de I-D+i 2008-2011, prorogued by agreement of the Consejo de Ministros of 2011/10/7.

References

[1] Dobra A, Lenkoski A, Rodriguez A. Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association* 2011; **106**(496):1418–1433.

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

[3] Wang F, Wall MM. Generalized common spatial factor model. *Biostatistics* 2003; **4**:569–582.

[4] Hogan JW, Tchernis R. Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. *Journal of the American Statistical Association* 2004; **99**:314–324.

[5] Tzala E, Best N. Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Statistical Methods in Medical Research* 2008; **17**(1):97–118.

[6] Marí-Dell'Olmo M, Martínez-Beneito MA, Borrell C, Zurriaga O, Nolasco A, Dominguez-Berjon MF. Bayesian factor analysis to calculate a deprivation index and its uncertainty. *Epidemiology* 2011; **22**(3):356–364.

[7] Knorr-Held L, Best N. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2001; **164**(13):73–85.

[8] Zhang Y, Hodges JS, Banerjee S. Smoothed ANOVA with spatial effects as a competitor to MCAR in multivariate spatial smoothing. *Annals of Applied Statistics* 2009; **3**(4):1805–1830.

[9] Mardia KV. Multidimensional multivariate Gaussian Markov random fields with application to image processing. *J. Multivariate Anal.* 1988; **24**(2):265–284, doi:10.1016/0047-259X(88)90040-1. URL `http://dx.doi.org/10.1016/0047-259X(88)90040-1`.

[10] Gelfand AE, Vounatsou P. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* 2003; **4**(1):11–25.

[11] Jin X, Carlin BP, Banerjee S. Generalized hierarchical multivariate CAR models for areal data. *Biometrics* 2005; **61**:950–961.

[12] Jin X, Banerjee S, Carlin BP. Order-free co-regionalized areal data models with application to multiple-disease mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2007; **69**(5):817–838.

[13] Macnab YC. On Gaussian Markov random fields and Bayesian disease mapping. *Statistical Methods in Medical Research* 2011; **20**:49–68.

[14] Martinez-Beneito MA. A general modelling framework for multivariate disease mapping. *Biometrika* 2013; .

[15] Lunn D, Thomas A, Best N, Spiegelhalter D. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337.

[16] Gentle JE. *Matrix Algebra. Theory, Computations, and Applications in Statistics.* Springer-Verlag, 2007.

[17] Rousseeuw PJ, Molenberghs G. The shape of correlation matrices. *The American Statistician* 1994; **48**:276–279.

[18] Harville DA. *Matrix algebra from a statistician's perspective.* Springer, 1997.

[19] Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathemathics* 1991; **43**:1–21.

[20] Stephens M. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2000; **62**(4):795–809.

[21] Zurriaga O, Martínez-Beneito MA, Botella-Rocamora P, López-Quílez A, Melchor I, Amador A, Vanaclocha H, Nolasco A. Spatio-temporal mortality atlas of Comunitat Valenciana 2010. URL `http://www.geeitema.org/AtlasET/index.jsp?idioma=I`, uRL: http://www.geeitema.org/AtlasET/atlas.jsp?idioma=I.

[22] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2009. ISBN 3-900051-07-0. URL: http://www.R-project.com.

[23] Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; **64**:583–641.

[24] Botella-Rocamora P, López-Quílez A, Martinez-Beneito MA. Spatial moving average risk smoothing. *Statistics in Medicine* 2012; doi:10.1002/sim.5704.

Model	Computing time **QR**-based model	Computing time **M**-based model
2	44.8	8.1
3	117.4	16.0
4	-	26.5
5	-	34.4
7	-	73.5
10	-	145.1
12	-	217.5
21	-	823.0

Table 1: Computation times, in minutes, for the **QR** and **M**-based model for different numbers of diseases.

Model	**DIC**
Factor model with 2 factors (proper CAR)	13941.49
Factor model with 3 factors (proper CAR)	13883.61
Factor model with 4 factors (proper CAR)	13786.78
Factor model with 5 factors (proper CAR)	13781.26
Factor model with 6 factors (proper CAR)	13772.47
Factor model with 7 factors (proper CAR)	**13766.58**
Factor model with 8 factors (proper CAR)	13767.62
Factor model with 9 factors (proper CAR)	13767.71
Factor model with 7 factors (BYM)	**13737.37**
Independent modelling (BYM)	**13890.08**

Table 2: DIC model selection criteria for different **M**-based models.

Model	DIC (dataset with 12 mortality causes)	DIC (dataset with 21 mortality causes)
Fixed effects model	24090.0	48282.4
Random effects model	24071.5	48282.8

Table 3: DICs for fixed and random effects models, for the studies including 12 and 21 diseases.

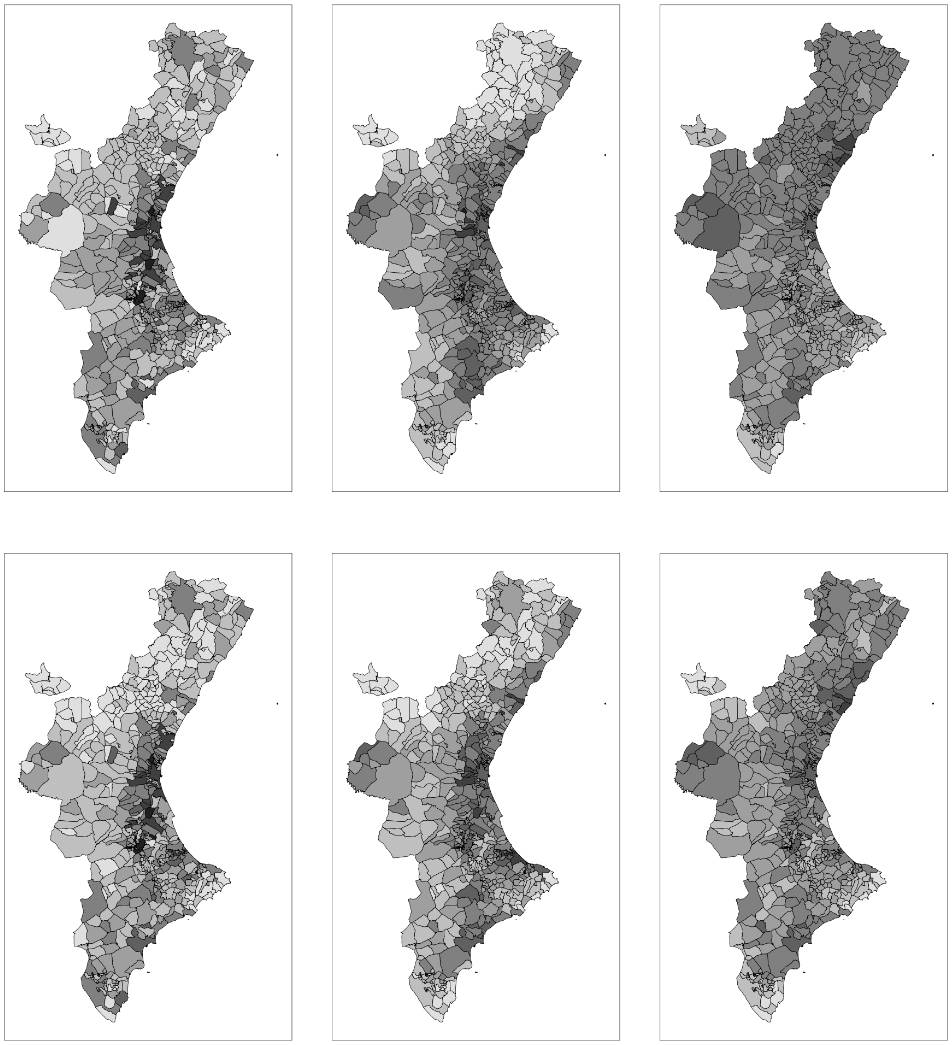

Figure 1: BYM independent modelling (upper row) and **M**-based BYM multivariate modelling (lower row) for Cirrhosis, Bladder Cancer and Rectum Cancer. Maps depict the posterior mean of relative risks for every municipality.

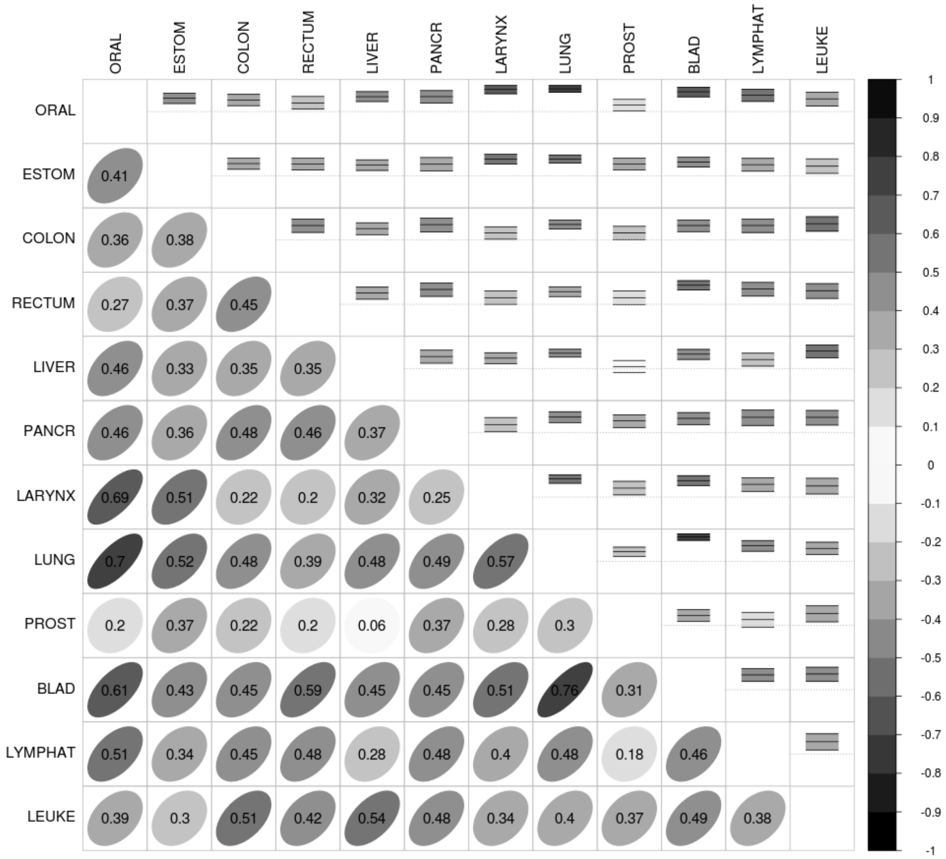

Figure 2: Correlations between geographical mortality patterns for the study considering just the cancer-related causes. Lower left-triangular side shows posterior means of those correlations. Upper right-triangular side shows 80% Posterior Credibility Intervals for that same correlations.