Published in final edited form as:

Stat Med. 2016 August 15; 35(18): 3189–3202. doi:10.1002/sim.6914.

The bivariate combined model for spatial data analysis

Thomas Neyens^{a,*}, Andrew B. Lawson^b, Russell S. Kirby^c, and Christel Faes^{a,d}

^aI-Biostat, Hasselt University, Hasselt, Belgium

^bDivision of Biostatistics and Epidemiology, College of Medicine, Medical University of South Carolina, Charleston, SC, USA

^cDepartment of Community and Family Health, College of Public Health, University of South Florida, Tampa, FL, USA

^dI-Biostat, KU Leuven - University of Leuven, Leuven, Belgium

Abstract

To describe the spatial distribution of diseases, a number of methods have been proposed to model relative risks within areas. Most models use Bayesian hierarchical methods, in which one models both spatially structured and unstructured extra-Poisson variance present in the data. For modeling a single disease, the conditional autoregressive (CAR) convolution model has been very popular. More recently, a combined model was proposed that 'combines' ideas from the CAR convolution model and the well-known Poisson-gamma model. The combined model was shown to be a good alternative to the CAR convolution model when there was a large amount of uncorrelated extra-variance in the data. Less solutions exist for modeling two diseases simultaneously or modeling a disease in two sub-populations simultaneously. Furthermore, existing models are typically based on the CAR convolution model. In this paper, a bivariate version of the combined model is proposed in which the unstructured heterogeneity term is split up into terms that are shared and terms that are specific to the disease or subpopulation, while spatial dependency is introduced via a univariate or multivariate Markov random field. The proposed method is illustrated by analysis of disease data in Georgia (USA) and Limburg (Belgium) and in a simulation study. We conclude that the bivariate combined model constitutes an interesting model when two diseases are possibly correlated. As the choice of the preferred model differs between data sets, we suggest to use the new and existing modeling approaches together and to choose the best model via goodness-of-fit statistics.

Keywords

Bivariate Modeling; Disease Mapping; Overdispersion; Correlated Gamma Random Effect; Multivariate Conditional Autoregressive Model

*Correspondence to: I-Biostat, Hasselt University, Hasselt, Belgium.

1. Introduction

In the last 20 years, the use of disease mapping has become well-established amongst epidemiological researchers. The growing amount of spatial information and the availability of easy-to-use software has facilitated the analysis of the geographical distribution of diseases. Concurrently, geographical clustering of diseases has become of increasing concern to the public, and the investigation of the spatial distribution of diseases has become a standard requirement by governmental agencies. Modern disease mapping methods allow one to investigate whether there are zones in an area with elevated risk, and can point to the existence of clusters of diseases throughout the area. Such clusters can point to an environmental exposure, not yet taken into account in the analysis. While disease mapping methodology has mainly focused on the geographical distribution of single diseases, there is also a growing attention to the description of the spatial distribution of multiple disease. Lawson [1] gave an overview of multivariate disease mapping models.

In this paper, we consider the situation of a study area subdivided into fixed spatial units (lattices), in which disease counts are observed in a fixed time period. Disease counts per area are available for different diseases or for different population groups, and interest is in the spatial distribution of both diseases or population groups, as well as in the correlation between the spatial distributions. Existing spatial modelling frameworks for multivariate data are based on the extension of the traditionally used Poisson-convolution model, assuming a Poisson distribution for the counts conditional on the spatial process, and assuming that the spatial process is the sum of a Gaussian Markov random field plus an additional unstructured Gaussian variation, such as proposed by Besag *et al.* [2], and illustrated in e.g. Clayton and Bernardinelli [3], Rue and Held [4] and Lawson [1] amongst others. The convolution model allows to account for both the overdispersion, also called uncorrelated heterogeneity (UH), via the unstructured variation, as well as for the spatial correlation, also called correlated heterogeneity (CH), via the Markov Random Field; explaining the wide use and applicability of the convolution model. Alternatively, as proposed by Neyens, Faes and Molenberghs [5], one can assume that the spatial process is the sum of a Gaussian Markov random field, to account for the spatial heterogeneity, plus a gamma-distributed unstructured heterogeneity term. This is called the combined model, which makes use of the strong conjugacy of the Poisson distribution with the gamma distribution. Details of the models are given in Section 3.

Extensions of the convolution model towards two or more diseases have been given by e.g. Knorr-Held and Best [6], Gelfand and Vounatsou [7] and Lawson [1] (chapter 10). Knorr-Held and Best [6] propose the use of a shared component model, assuming a shared Gaussian Markov Random Field for both diseases. Gelfand and Vounatsou [7] propose a multivariate conditional autoregressive model, introducing correlation in the spatial component. Lawson [1] discusses introducing correlation in both the aspatial component, and in the spatially structured component. This has been used also by Kramer and Williamson [8], showing the flexibility of the latter model and the possibility to quantify the correlation between the spatial processes. In this paper, an alternative extension of the latter model is proposed, with correlation between the aspatial components introduced via gamma-distributed random effects. This forms an alternative method to the commonly assumed

Gaussian convolution model, providing the practitioner with more tools to efficiently model the bivariate disease distributions.

In Section 2, two case studies are introduced: the study of asthma and COPD (chronic obstructive pulmonary disease) in Georgia (USA), and the study of bladder cancer in males and females in Limburg (Belgium). Section 3 reviews the general framework of the CAR convolution and the combined model. Section 4 gives an explanation on existing methods to model spatial counts bivariately and proposes a new bivariate extension of the combined model. The data application is covered in Section 5, while a simulation study is presented in Section 6 and a conclusion is provided in Section 7.

2. Case Studies

In this section, two examples are given in which interest is in the bivariate spatial distribution of disease cases. The first case study focuses on the association between two diseases with similar aetiology, asthma and COPD (chronic obstructive pulmonary disease) in Georgia (Section 2.1). The second case study focuses on the correlation of the disease risk of bladder cancer between different groups, namely males and females (Section 2.2).

2.1. Asthma and COPD in Georgia

The first dataset represents counts of new cases of asthma and COPD in all 159 counties of Georgia (USA) in 2005. These respiratory diseases are likely to show spatial or non-spatial correlations since it is expected that they may have common etiological factors or determinants. The two diseases have similar mean numbers and standard deviations, while maximum asthma values are almost twice as high as the COPD counts (Table 1). Although the largest population sizes are found in the northwest around Atlanta, SIR (standardized incidence rate) estimates are highest in the southeastern region (Figure 1). This data set was also analysed by Lawson [1].

2.2. Bladder cancer in Limburg

The second dataset is part of the Limburgs Cancer Registry (LIKAR) in which all cancers in the province of Limburg (Belgium) are registered with additional information about each cancer type per region, age and gender between the years 1996 and 2005 [9]. The Limburg Cancer Registry focusses on the province Limburg, situated in the northeast of Belgium. With a total of 44 municipalities and the most populated areas centered in the middle of the province, Limburg is considered as the least urbanized province (350 inhabitants/km²) in the upper part of Belgium (Flanders). In this study, bladder cancer counts in males and females were investigated, with interest in the association between the two population groups. Summary statistics are provided in Table 1. This cancer type is diagnosed relatively frequently and it is clear that the southwestern towns have increased SIR's, for both males and females (Figure 2).

3. Univariate Disease Mapping

Assume there are $L = 2$ diseases or population groups. In this section, a separate analysis for every disease or group is considered. Assume that the study region is divided into N sub-

areas, and define the disease counts $\{y_{il}\}$ for disease or groups $l = 1, \dots, L$ within spatial unit $i = 1, \dots, N$. The corresponding expected number of disease cases, by applying e.g. an internal age-standardisation, are denoted by $\{e_{il}\}$. The unknown relative risks are given by $\{\theta_{il}\}$. A basic Poisson model assumes that $\{y_{il}\}$ follows a Poisson distribution with mean parameter $e_{il}\theta_{il}$, with all the θ_{il} independent:

$$y_{il} \sim \text{Poisson}(e_{il}\theta_{il}). \quad (1)$$

This leads to crude risk estimates (SIRs) of $\theta_{il} = y_{il}/e_{il}$. As mentioned by several authors, this model is insufficient, because of the heterogeneity and spatial association in real-life applications. A first extension of this basic model, by allowing for overdispersion, results in the so-called Poisson-gamma and Poisson-lognormal models. The Poisson-gamma model is defined by (1) together with

$$\theta_{il} \sim \Gamma(a_l, b_l), \quad (2)$$

with $\Gamma(a_l, b_l)$ a gamma distribution with scale parameter a_l and rate parameter b_l . The relative risks θ_{il} have prior mean a_l/b_l and variance a_l/b_l^2 . This model is attractive as a result of the conjugacy of the Poisson and gamma distribution, allowing easy analytical expression

for this model. When assuming a_l and b_l to be fixed, the posterior mean of θ_{il} is $\frac{a_l + y_{il}}{b_l + e_{il}}$, which can be re-written as a weighted average of y_{il}/e_{il} and a_l/b_l . Note that the prior predictive mean for y_{il}/e_{il} is equal to $a_l/b_l \equiv \mu_l$, the same as the prior mean for θ_{il} , but its variance is $\mu_l \left(\frac{1}{e_{il}} + \frac{1}{b_l} \right)$.

The Poisson-lognormal model on the other hand is defined by replacing (2) with

$$\log(\theta_{il}) = \alpha_l + v_{il} \quad (3)$$

$$v_{il} \sim N(0, \sigma_{v_l}^2). \quad (4)$$

Analytical deviations of the posterior distribution for θ are less straightforward in this situation, although the prior predictive distribution (or equivalently, the marginal likelihood) for y_{il}/e_{il} can be derived in a similar way as in Molenberghs, Verbeke and Demétrio [10],

which has mean given by $\exp(\alpha_l + \frac{1}{2}\sigma_{v_l}^2) \equiv \mu_l$ and variance equal to $\mu_l/e_l + \mu_l^2(\exp(\sigma_{v_l}^2) - 1)$. From this it is clear that the mean-variance relationship differs from the one in the Poisson-gamma model, resulting in differences in the way smoothing is performed by the two models.

Both models can further be extended by the addition of spatial correlation, resulting in so-called spatial models, for instance by the inclusion of a conditional autoregressive (CAR) random effect [2]. The well-known convolution model for spatial data is defined by (1) together with

$$\log(\theta_{il}) = (\alpha_l + \mathbf{x}_i' \boldsymbol{\beta}_l + u_{il} + v_{il})$$

where u_{il} and v_{il} represent respectively spatial and aspatial heterogeneity for the i th area. The aspatial random effect v_{il} is defined in (5). The spatial random effects term u_{il} is assumed to follow an intrinsic *Conditional Autoregressive (CAR)* prior such as introduced by Besag and Kooperberg [11]:

$$u_i | u_{j, i \neq j} \sim N(\bar{\mu}_i, \sigma_i^2), \quad (5)$$

with $\bar{\mu}_i = \sum_{j=1}^N w_{ij} u_j / \sum_{j=1}^N w_{ij}$, $\sigma_i^2 = \sigma_u^2 / \sum_{j=1}^N w_{ij}$ and $w_{ij} = 1$ if areas i and j are adjacent and 0 otherwise. This method of weighting is just one out of many though, with Bivand, Pebesma and Gómez-Rubio [12] providing a range of different weighting schemes.

The combined model, on the other hand, is defined by (1) together with

$$\theta_{il} = g_{il} \exp(\alpha_l + \mathbf{x}_i' \boldsymbol{\beta}_l + u_{il})$$

with u_{il} as in (5) and $g_{il} \sim \Gamma(a_l, b_l)$, as in the Poisson-gamma model. An important strength of this adaptation is the closed-form conditional distribution due to the strong conjugacy [13, 14] between the Poisson and gamma distribution. Conjugacy means, without going into details, that the hierarchical and random-effects densities are algebraically alike. Strong conjugacy is a property reflecting that conjugacy holds in the presence of the random effects [10]. Derivation of the posterior mean of y_i conditional on the spatial random effect u_i is

again easy, and becomes $\frac{a_l + y_{il}}{b_l + \kappa_{il} e_{il}}$ with $\kappa_{il} = a_l + \mathbf{x}_i' \boldsymbol{\beta}_l + u_{il}$, which can be written as a weighted average of y_{il}/e_{il} and a_l/b_l . It should be clear that a lot of shrinkage to the prior mean a/b will occur for rare diseases and small areas, which is similar to the Poisson-gamma model. But in contrast to the Poisson-gamma model, the weights w_i depend on the spatial structure κ_i , in other words, the amount of smoothing is also spatially structured.

When the amount of uncorrelated heterogeneity (UH) in the data is large, the combined model has been shown by simulations studies to outperform the traditional CAR convolution model in terms of DIC (deviance information criterion) and MSPE (mean squared predictive error). In fact, although the CAR convolution model's robustness when simulating a wide range of underlying true risk models has made it popular in spatial disease mapping [15], it typically models a lot of extra-variability as correlated heterogeneity. In a number of cases, this can in fact become problematic, since an oversmoothed RR map can underestimate the effect of non-spatial factors. On the other hand, when the extra-variance is mainly induced

by spatial factors, the combined model wrongfully addresses a (too) large amount of it to an unstructured process. [5].

4. Bivariate Disease Mapping

Although the univariate disease map gives the researcher a number of answers to scientific questions, issues concerning interactions between the two diseases remain unsolved and require a bivariate approach.

4.1. Bivariate convolution model

To put the definitions of the proposed bivariate combined model into perspective, it is important to keep in mind that basically two strategies exist in the bivariate setting, namely models with common random effects and models with correlated random effects. The first type of models assumes that specific spatial or non-spatial extra-variance terms are shared between the models, while both models may also have a set of separate terms. As an extension of the convolution model, one could consider the following model (for an overview of models, Lawson [1]):

$$\theta_{1i} = \exp(\alpha_1 + u_{1i} + v_i),$$

$$\theta_{2i} = \exp(\alpha_2 + u_{2i} + v_i),$$

with u_{1i} and u_{2i} defined univariately as mentioned in (5) and $v_i \sim \mathcal{N}(0, \sigma_v^2)$. Alternatively, one can choose to take the spatial random effects as shared ($u_{1i} = u_{2i} = u_i$) while the uncorrelated heterogeneity (UH) term may be taken as disease-specific (v_{ih} , $h = 1, 2$). When the diseases are not equally common, it is proposed to take into account a scaling component [6]. By doing so, the amount of overdispersion explained by the random effects is not equal anymore. As an example, a typical model is:

$$\theta_{1i} = \exp(\alpha_1 + u_{1i} + \delta v_i),$$

$$\theta_{2i} = \exp(\alpha_2 + u_{2i} + v_i/\delta),$$

with δ the scaling component. On the other hand, it is not difficult to understand that models with shared terms are somehow too restrictive and indeed, intuitively two related diseases are seen as diseases that act alike, not the same. This can be modeled by assuming that both random effects are correlated. In the disease mapping context, such models are not yet widely available, although the multivariate CAR convolution model has been developed [16], which in the bivariate framework uses the MCAR specification, which models spatial correlation on a multivariate scale [7], while it uses $(v_{1i}, v_{2i}) \sim MV\mathcal{N}(0, \Sigma)$ for the aspatial extra-variance. As an alternative to this model, we propose a bivariate extension of the combined model, given in the next section.

4.2. Bivariate combined model

In the bivariate combined model, we assume that

$$y_{i1} \sim \text{Poisson}(e_{i1}\theta_{i1}),$$

$$y_{i2} \sim \text{Poisson}(e_{i2}\theta_{i2}),$$

with the relative risks modeled as

$$\theta_{i1} = g_{i1} \exp(\alpha_1 + u_{i1}),$$

$$\theta_{i2} = g_{i2} \exp(\alpha_2 + u_{i2}).$$

The random effects u_{1j} and u_{2j} are conditional autoregressive random effects, and are specified either univariately as in (5) or multivariately, using the MCAR specification [7, 16]. The gamma-distributed overdispersion random effects g_{i1} and g_{i2} can also be modeled either uni- or bivariate, with the bivariate specification defined as follows:

$$g_{i1} = \frac{1}{k_0 + k_1} (k_0 \gamma_{i0} + k_1 \gamma_{i1})$$

$$g_{i2} = \frac{1}{k_0 + k_2} (k_0 \gamma_{i0} + k_2 \gamma_{i2}).$$

where k_0 , k_1 and k_2 are three real positive variables. The terms γ_0 , γ_1 , and γ_2 are assumed to be independent gamma-distributed random variables. The common term is assumed to be $\gamma_0 \sim \Gamma(1, 1)$, and the two disease- or population-specific terms are $\gamma_1 \sim \Gamma(1, 1)$ and $\gamma_2 \sim \Gamma(1, 1)$,

such that $E(g_{i1}) = 1$ and $E(g_{i2}) = 1$, while $\text{Var}(g_{i1}) = \frac{k_0^2 + k_1^2}{(k_0 + k_1)^2}$ and $\text{Var}(g_{i2}) = \frac{k_0^2 + k_2^2}{(k_0 + k_2)^2}$. Note that k_0 can be interpreted as a factor that denotes the amount of shared non-spatial variability between both outcomes, while k_1 and k_2 have values that represent the amount of non-spatial variability that is specific to each outcome. In the most complex but least restrictive setting, where u_{1j} and u_{2j} are MCAR distributed and g_{1j} and g_{2j} are bivariate gamma distributed, one is able to capture both spatial and non-spatial correlated variability in the data. According to the particular situation, more restrictions can be added, such as e.g. the use of two disease-specific univariate CAR random effects instead of a MCAR specification, or even a model with only the bivariate g_{1j} and g_{2j} , which may suffice if there is no spatial correlation at play. Indeed, many options are available and depending on the context, some models will fare better than others, as is depicted in the following case and simulation studies.

5. Data Application

In this section, we compare the different models within this modeling framework. It can be seen that we have a large set of possible models, amongst which the best model can be chosen. Several models will be considered for both case studies and are listed in Table 2. Included are univariate models fitted separately on the two outcomes, joint models with a correlated random effect and joint convolution models (two random effects).

Goodness-of-fit (GOF) within Bayesian inference has always caused a number of issues, since using complex forms of hierarchy is known for causing GOF statistics to be unstable. Although a number of possibilities have been proposed, most studies use the well-known *deviance information criterion (DIC)* [17]. On one hand DIC was used, but with pD , the effective number of parameters, as an extra statistic was used in this study. A negative value for pD could indicate the need to use appropriate adjusted GOF statistics [18] while looking at pD could help in making the final decision DIC values, when DIC values only differed slightly between models. Secondly, we used the *mean squared predictive error (MSPE)* to investigate the overall loss across the data. MSPE is an average of the item-wise squared error loss, $MSPE = \sum_i \sum_j (y_i - y_{ij}^{pr})^2 / (G \times m)$, with y_{ij}^{pr} being the predictive data item at MCMC sampler's iteration j , m being the number of observations and G being the sampler sample size. It is important to note that these measures are both GOF statistics in their own right, but since the DIC measures the global goodness-of-fit and MSPE is a measure for predictive ability, they do not provide the same information. Furthermore, the empirically-based Pearson correlations between both diseases/population groups were calculated for the estimated spatial random effects, non-spatial random effects and relative risks, i.e.

$r_{x_1, x_2} = \frac{\text{cov}(x_1, x_2)}{sd(x_1)sd(x_2)}$ with $x = u_i, v_i, g_i$ or θ_i . Note that a burn-in of 10000 iterations was used with an additional 20000 iterations to achieve convergence.

5.1. Asthma and COPD in Georgia

Table 3 gives an overview of the model fits. For the Georgia data, the best model according to DIC was achieved by model 3 (d), with an MCAR CH term and a bivariate gamma UH term. It is also apparent that all models with gamma bivariate terms had consistently smaller DIC values than those with normal bivariate UH random effects, e.g. when comparing bivariate model 2 (b) with 2 (c) or the convolution bivariate models 3 (a) with 3 (c) and 3(d). Further, it can be observed that the univariate models do not seem to fit much worse than bivariate models, e.g. univariate model 1 (a) had a lower DIC value than many other bivariate models, such as the convolution model 3 (b). In terms of MSPE, different things can be seen: Model 2 (c), the model with only a bivariate gamma UH term, had the lowest error. Although differences between the bivariate gamma and other models, such as 1 (a), 2 (b) or 3 (c) were small, it can be stated that also in terms of MSPE, the bivariate gamma term performed equally or better than the bivariate normal term. Lastly, it is interesting to point out that of all models, model 2 (a), with only a MCAR term, did worse in terms of both DIC and MSPE. This may be due to the combination of the need for a UH random effects term and the non-necessity of a MCAR term as spatial random effect.

The last column in Table 3 shows the empirically-based correlation estimates for the relative risks. These estimates summarize the correlation between COPD and asthma. From all models, we can see that the correlation between the two diseases is estimated around 0.5. At first sight, this is unexpected, since the first two models (1 (a) and 1 (b)) model the two diseases independently. This however indicates that these models are indeed insufficient in understanding the source of association between the diseases. In addition, when modelling the association between the two diseases, correlations increase to around 0.6, indicating that a misspecification of the models might underestimate the correlation. Among the bivariate models, the correlation of the RR's remains relatively alike. However, in order to understand the source of association, also the empirical correlations of the random effect terms were calculated. Looking at the best fitting models (model 3 (d) based on DIC and model 2 (c) based on MSPE), a significant correlation is observed for the non-spatial heterogeneity terms and a non-significant correlation for the spatial component. This consistently indicates that the specific spatial structure as taken into account by the CH terms in this study does not add to the correlation between asthma and COPD.

When looking at the relative risk maps (Figure 3), based on model 3 (d), we indeed see that both diseases are correlated, with a shared increase in relative risk in the central southeastern part of Georgia, while there are also disease-specific patterns scattered around the map. When investigating Figure 4, which shows maps of the bivariate gamma-distributed UH random effects, we see that both diseases have elevated UH values in southeastern Georgia. But indeed, Figure 4 only gives information about the UH in the data, while the RR's on Figure 3 also include the correlated heterogeneity (CH), which is shown in Figure 5. Indeed, also the CH maps show higher values in southeastern Georgia, which suggests that in this case extra-variance, which was not taken into account by the spatial random effect, was modeled by the bivariate gamma random effects g_{11} and g_{12} .

5.2. Bladder cancer in Limburg

Results for the Limburg data are summarized in the bottom panel of Table 3. Again, the convolution models, which combine CH with UH terms, did not have better fits than the other bivariate models (models 2 (a) – 2 (c)). In fact, the model with the lowest DIC was model 2 (c), which only has a bivariate gamma term. Furthermore, it again is striking that the bivariate gamma term provides better fits than the bivariate normal one, e.g. when comparing DIC values from model 2 (b) with 2 (c), or the convolution models 3 (a) with 3 (c), the models with a bivariate gamma term consistently turn out as the better ones. Model 3 (d) however, which uses MCAR and bivariate gamma terms did not show an improved fit compared to model 3 (b). This is probably due to the model being too complex for these particular data with a small sample size. When looking at the MSPE statistics, it seems that models without UH terms (Models 1 (b) – 2 (a)) do worse than the others. The best models in terms of MSPE are models 3 (c) (MCAR CH random effects and a bivariate gamma random effect) and 2 (c) (only a bivariate gamma term). When looking at the DIC and MSPE investigations together, with an emphasis on model simplicity rather than complexity, model 2 (c) would then constitute as the best model, though differences are small.

When looking at the RR correlation estimates, we see that model 1 (a) has a substantially smaller estimate than the other models. This is due to the univariate nature of these first two models. Furthermore, RR correlation estimates are less stable among the models when compared with those given by the Georgia data analyses, with empirical correlation estimates ranging between 0.3514 and 0.7400. For model 2 (c), which was previously chosen as the best fitting model, the estimated RR correlation was 0.7076. When focussing on the random effects correlation estimates, again relatively high values are estimated for the univariate models, but these are likely to underestimate the true correlation. Again all correlation estimates are positive, but now not all of them differ significantly from zero. When looking at the convolution models in particular, the spatial random effects correlations are never significant and for model 3 (b) also the non-spatial random effects are not significantly correlated between males and females. However, RR correlation estimates do show significant positive correlations for all models. When looking at the RR estimates (Figure 6) and the bivariate gamma random effects for the uncorrelated heterogeneity, based on the preferred model 2 (c), it is clear that there is a high correlation between the male and female cases with elevated risks in the southwestern parts of Limburg. Note that these RR's were estimated via a non-spatial bivariate combined model, which is strange when looking at Figures 6 and 7 as they seem to show spatial variability. However, in this case where the model may have problems in distinguishing between different forms of variability, the gamma random effects may capture a part of the spatial correlation.

6. Simulation Study

A simulation study was set up to investigate the different models when knowing the underlying model. Based on the bladder cancer case study in Limburg, data were generated for 44 municipalities, with $E_i = 1$ ($\forall i = 1, \dots, 44$) from 2 processes, namely (1) a MCAR and bivariate gamma convolution model (Model 3(d)) and (2) a bivariate gamma model (Model 2(c)). The MCAR random effects were generated using the adjacency structure in Limburg. As true values, we chose $\alpha_1 = \alpha_2 = 2$, $k_0 = 10$, $k_1 = k_2 = 8$, $Var(u_{1i}) = Var(u_{2i}) = 3$ and $Cov(u_{1i}, u_{2i}) = 2.9$. Note that by doing this, relatively strong correlations were simulated between both spatial and non-spatial random effects. For both underlying processes, 100 datasets were generated, which were analysed with all models 1(a) to 3(d), which were previously used in the case studies (Table 2). Mean DIC and MSPE values were calculated to evaluate goodness-of-fit among the models, while mean squared error

($MSE = \sum_{j=1}^N (\hat{\theta}_j - \theta^{\text{true}})^2 / N$, with N the number of simulated data sets) was used to evaluate parameter estimation.

For the first underlying process, namely the one based on a MCAR and bivariate gamma model (upper panels of Table 4), DIC values mostly favour model 3(d) (MCAR and bivariate gamma). In terms of MSPE, 2(b) is mostly preferred. Note that the true underlying model, 3(d), has consistently higher MSPE values than the other convolution models. While this might come as a surprise, it is not unexpected that the predictive performance of a less restrictive model (e.g. Model 2(d)) performs better than the more restrictive model, such as the true model 3(d). Furthermore, correlation between spatial random effects is underestimated by all models, although 3(b) and 3(d) come close to estimating the true

values. In contrast, non-spatial correlation is overestimated by all models, except for 3(b). It is also interesting to see that the total RR correlation estimates practically coincide for all convolution models. This becomes clear when looking at 3(c), where the correlation between spatial random effects is severely underestimated. This lack in explained correlation is however compensated through the overestimation of the non-spatial RE correlation, which yields estimations of RR correlations close to the estimations given by models that correctly account for the spatial and non-spatial correlations. The issue of overestimating UH when it actually comes from a spatial process, was also seen in the Limburg case study. It is likely that the sample size ($N=44$) also plays a role here, as it becomes difficult for the model to distinguish between different forms of extra-variability.

For the bivariate gamma underlying process (Process 2) (lower panels of Table 4), DIC mostly favours the true underlying model 2(c) and model 3(c). MSPE values however are very alike between all models, making it difficult to choose a best model based on that GOF statistic. In terms of random effects correlations, it is noteworthy to show that the MCAR random effects wrongfully estimate a correlation of 0.30–0.35 (models 3(b) and 3(d)), while correlations between non-spatial random effects are overestimated too for almost all models. This results in a general overestimation of the RR correlations for all bivariate models, except 3(b), for which it is slightly underestimated. From both simulation settings, it is difficult to conclude which model will behave best in certain circumstances. The bivariate gamma model and the MCAR + bivariate gamma model do a good job at explaining different sources of association, and are interesting to compare with other existing methods. Based on different GOF statistics, the practitioner can decide which model to finally opt for.

7. Concluding Remarks

In this paper, a novel method for the bivariate analysis of spatial disease counts is proposed, and an investigation of the source correlation between diseases or populations is presented. As indicated by the modeling results in the case studies and simulation study, the bivariate combined model, which introduces a bivariate gamma distributed random effects term to capture non-spatial extra-variance, fits better than a model with the same spatial random effects but with a multivariate normal random effect for UH. However, as data sets differ, the best way to analyse them will differ too. The bivariate combined model does not provide the best modeling approach in all cases and it is not straight-forward to predict when it does. It is likely that limited sample sizes may hamper estimation in complex models, but this is a well-known issue in statistics. We suggest the practitioner to investigate the results given by all models and to use GOF statistics to choose the best option. We conclude that the bivariate combined model proposes an interesting option when working with two possibly related datasets existing out of counts, and as such it has been shown that the bivariate combined model presents an interesting new piece in the spatial statistician's toolbox.

Acknowledgments

Financial support from the IAP research network #P6/03 of the Belgian Government (Belgian Science Policy) and the Research Foundation Flanders is gratefully acknowledged. The authors also acknowledge support from NIH R01CA172805.

References

1. Lawson, AB. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. 2nd. Boca Rotan: Chapman & Hall; 2008.
2. Besag J, York J, Mollié A. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*. 1991; 43:1–59.
3. Clayton, DG.; Bernardinelli, C. Bayesian methods for mapping disease risk. In: Elliott, P.; Cuzick, J.; English, D.; Stern, R., editors. *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*. Oxford: Oxford University Press; 1992.
4. Rue, H.; Held, L. *Gaussian Markov Random Fields: Theory and Applications*. New York: Chapman & Hall/CRC; 2005.
5. Neyens T, Faes C, Molenberghs G. A Generalized Poisson-Gamma Model for Spatially Overdispersed Data. *Spatial and Spatio-temporal Epidemiology*. 2012; 3:185–194. [PubMed: 22749204]
6. Knorr-Held L, Best NG. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society*. 2001; 164:73–85.
7. Gelfand A, Vounatsou P. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*. 2003; 4:11–25. [PubMed: 12925327]
8. Kramer MR, Williamson R. Multivariate Bayesian spatial model of preterm birth and cardiovascular disease among Georgia women: Evidence for life course social determinants of health. *Spatial and Spatio-temporal Epidemiology*. 2013; 6:25–35. [PubMed: 23973178]
9. <http://likas.edm.uhasselt.be/>.
10. Molenberghs G, Verbeke G, Demétrio CGB, Vieira AMC. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical science*. 2010; 25:325–347.
11. Besag J, Kooperberg C. On conditional and intrinsic autoregressions. *Biometrika*. 1995; 82:733–746.
12. Bivand, RS.; Pebesma, EJ.; Gómez-Rubio, V. *Applied Spatial Data Analysis with R*. New York: Springer; 2008.
13. Cox, DR.; Hinkley, DV. *Theoretical Statistics*. London: Chapman & Hall; 1974. p. 370
14. Lee, Y.; Nelder, J.; Pawitan, Y. *Generalized Linear Models With Random Effect*. London: Chapman & Hall; 2006. p. 178
15. Lawson AB, Biggeri AB, Boehning D, Lesaffre E, Viel JF, Clark A, Schlattmann P, Divino F. Disease mapping models: an empirical evaluation. *Statistics in Medicine*. 2000; 19:2217–2241. [PubMed: 10960849]
16. Stern, HS.; Cressie, N. Inference for extremes in disease mapping. In: Lawson, A.; Biggeri, A.; Bohning, D.; Lesaffre, E.; Viel, JF.; Bertollini, R., editors. *Disease Mapping and Risk Assessment for Public Health*. Chichester: John Wiley and Sons; 1999. p. 63-84.
17. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian deviance, the effective number of parameters and the comparison of arbitrarily complex models. *Journal of the Royal Statistical Society B*. 2002; 64:583–640.
18. Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, D. *Bayesian Data Analysis*. London: Chapman & Hall/CRC press; 2004.

Figure 1.
Standardized Incidence Rates ($SIR_i = O_i/E_i$) per county for asthma (left panel) and COPD (right panel) in Georgia.

Figure 2.
Standardized Incidence Rates for bladder cancer ($SIR_i = O_i/E_i$) per municipality for males (left panel) and females (right panel).

Figure 3.
RR maps for asthma and COPD counts in Georgia, based on a model with 2 MCAR random effects and 2 bivariate gamma random effects (model 3 (d)).

Figure 4.

Maps of g_{11} and g_{22} for asthma and COPD counts in Georgia, based on a model with 2 MCAR random effects and 2 bivariate gamma random effects (model 3 (d)).

Figure 5.

Disease-specific CH maps for asthma and COPD counts in Georgia, based on a model with 2 MCAR random effects and 2 bivariate gamma random effects (model 3 (d)).

Figure 6.
RR maps for male and female bladder cancer counts in Limburg, based on a model with the bivariate gamma random effects term (model 2 (c)).

Figure 7.

Maps of g_{i1} and g_{i2} for male and female bladder cancer counts in Limburg, based on a model with the bivariate gamma random effects term (model 2 (c)).

Table 1

Summary statistics for two case studies: (1) asthma and COPD counts in the counties of Georgia (left); and (2) male and female bladder cancer in the municipalities of Limburg (right).

	Georgia		Limburg	
	Asthma	COPD	Bladder Cancer Male	Female
Mean	72.15	92.99	23.82	5.46
Standard deviation	138.48	112.93	29.85	8.21
Minimum	0	0	0	0
Maximum	1105	697	192	48

Table 2

Case study models (“-” indicates that the statistic does not exist for the specific model)

Model	Family	Random Effects	
		Spatial	Non-Spatial
1(a)	Univariate models on two outcomes	-	disease-specific univariate gamma
1(b)	Bivariate models with correlated random effects	disease-specific UCAR	-
2(a)	Bivariate models with correlated random effects	MCAR	-
2(b)	Bivariate models with correlated random effects	-	bivariate normal
2(c)	Bivariate models with correlated random effects	-	bivariate gamma
3(a)	Bivariate convolution models	disease-specific UCAR	bivariate normal
3(b)	Bivariate convolution models	MCAR	bivariate normal
3(c)	Bivariate convolution models	disease-specific UCAR	bivariate gamma
3(d)	Bivariate convolution models	MCAR	bivariate gamma

Table 3

Model fits and empirically-based correlations (with 95% credible intervals) between random effects and relative risks (“-” indicates that the statistic does not exist for the specific model).

Model	Random Effects		Model Fit		EB RE Corr		EB	
	Spat.	Non-Spat.	DIC	pD	MSPE	Spatial	Non-spatial	RR corr
Georgia Data								
1(a)	-	univ. gam. ^a	2378.5	281.2	330.7	-	0.5003 [0.4305; 0.5682]	0.5003 [0.4305; 0.5682]
1(b)	UCAR ^b	-	2394.7	274.1	347.5	0.4699 [0.4050; 0.5325]	-	0.5184 [0.4511; 0.5832]
2(a)	MCAR ^c	-	2469.2	306.0	410.9	0.5735 [0.5071; 0.6354]	-	0.6104 [0.5434; 0.6734]
2(b)	-	biv. norm. ^d	2385.3	274.4	330.4	-	0.5633 [0.4941; 0.6293]	0.5929 [0.5229; 0.6595]
2(c)	-	biv. gam. ^e	2355.7	272.3	329.2	-	0.6205 [0.5501; 0.6870]	0.6205 [0.5501; 0.6870]
3(a)	UCAR	biv. norm.	2377.2	272.1	336.4	0.0719 [-0.3975; 0.5890]	0.6257 [0.5024; 0.7336]	0.5839 [0.5133; 0.6498]
3(b)	MCAR	biv. norm.	2396.4	282.1	345.1	0.6302 [0.2310; 0.8264]	0.3951 [0.0114; 0.7114]	0.5873 [0.5181; 0.6521]
3(c)	UCAR	biv. gam.	2351.1	268.7	330.9	0.0232 [-0.4724; 0.5117]	0.6525 [0.5667; 0.7433]	0.6251 [0.5554; 0.6908]
3(d)	MCAR	biv. gam.	2344.3	255.5	346.5	0.4584 [-0.0028; 0.7649]	0.6790 [0.5537; 0.7908]	0.6130 [0.5439; 0.6761]
Limburg Data								
1(a)	-	univ. gam.	465.0	54.9	57.8	-	0.3514 [0.1200; 0.5590]	0.3514 [0.1200; 0.5590]
1(b)	UCAR	-	459.8	43.8	62.9	0.4765 [0.2124; 0.6858]	-	0.5032 [0.2306; 0.7038]
2(a)	MCAR	-	459.6	44.0	62.8	0.7339 [0.4854; 0.8865]	-	0.7400 [0.5071; 0.8861]
2(b)	-	biv. norm.	459.8	51.1	56.9	-	0.5915 [0.3171; 0.7933]	0.6118 [0.3637; 0.7985]
2(c)	-	biv. gam.	457.5	52.8	56.4	-	0.7076 [0.4628; 0.8959]	0.7076 [0.4628; 0.8959]
3(a)	UCAR	biv. norm.	460.5	51.7	57.4	0.0305 [-0.4455; 0.5015]	0.5644 [0.2449; 0.7822]	0.5875 [0.3221; 0.7858]
3(b)	MCAR	biv. norm.	463.0	56.1	56.8	0.5221 [-0.0621; 0.8461]	0.3094 [-0.1824; 0.6904]	0.5294 [0.2685; 0.7355]
3(c)	UCAR	biv. gam.	457.9	53.1	56.2	0.0010 [-0.4488; 0.4604]	0.7033 [0.4516; 0.8922]	0.7001 [0.4512; 0.8885]
3(d)	MCAR	biv. gam.	468.1	60.3	60.3	0.3689 [-0.3498; 0.8282]	0.6554 [0.3053; 0.9016]	0.6034 [0.3494; 0.7997]

^aTwo disease-specific univariate gamma random effects.

^bTwo disease-specific univariate conditionally autoregressive random effects.

^cTwo multivariate conditionally autoregressive random effects.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Two bivariate normal random effects.
 $T_{p,2}$
Two bivariate gamma random effects.

Table 4

Simulation study results for both underlying processes, namely (1) the MCAR + bivariate gamma convolution model and (2) the bivariate gamma model. Mean empirical correlations with standard deviations between parentheses and their MSE are given for the relative risks, the spatial random effects and the non-spatial random effects. For DIC and MSPE, mean values are given. Between parentheses, the number of simulated data sets is indicated for which a particular model had the lowest DIC (MSPE) or a DIC (MSPE) value less than 5 (2) units from the lowest DIC (MSPE) within the same simulation run. Note that the empirically based random effects and relative risk correlation values for the true models were calculated as the average of the true empirically based correlations over 100 simulated data sets. Best DIC, MSPE and MSE values are indicated in bold. Lastly, “-” indicates that the statistic does not exist for the specific model.

Process 1: MCAR + bivariate gamma									
Model	Random Effects		Model Fit		EB RE Corr		EB		RR corr
	Spat.	Non-Spat.	DIC	MSPE	Spatial	Non-spatial	Spatial	Non-spatial	
True	MCAR ^c	biv. gam ^e	-	-	0.9802	0.5985			0.8101
1(a)	-	univ. gam. ^a	512.8 (0)	76.4 (73)	-	0.7731 (0.0115)			0.7731 (0.0115)
1(b)	UCAR ^b	-	529.9 (0)	99.3 (0)	0.7963 (0.0046)	-			0.7788 (0.0119)
2(a)	MCAR	-	530.2 (0)	108.7 (0)	0.8961 (0.0065)	-			0.8238 (0.0125)
2(b)	-	biv. norm. ^d	497.6 (1)	74.8 (100)	-	0.8795 (0.0063)			0.8127 (0.0119)
2(c)	-	biv. gam.	495.2 (16)	88.4 (0)	-	0.8400 (0.1559)			0.8400 (0.1559)
3(a)	UCAR	biv. norm.	499.3 (1)	76.4 (85)	0.0522 (0.0376)	0.8711 (0.0385)			0.8094 (0.0122)
3(b)	MCAR	biv. norm.	506.2 (0)	84.0 (7)	0.9014 (0.0535)	0.5974 (0.0438)			0.8023 (0.0118)
3(c)	UCAR	biv. gam.	496.1 (15)	87.0 (1)	0.4660 (0.0851)	0.9301 (0.0352)			0.8110 (0.0124)
3(d)	MCAR	biv. gam.	481.0 (88)	90.9 (0)	0.9009 (0.0191)	0.8065 (0.0373)			0.8111 (0.0121)

Model	Random Effects		MSE RR		MSE EB RE Corr		MSE EB	
	Spat.	Non-Spat.	RR1	RR2	Spatial	Non-spatial	RR corr	RR corr
1(a)	-	univ. gam.	948.2	1420.3	-	0.0577		0.0016
1(b)	UCAR	-	968.6	1453.0	0.0350	-		0.0011
2(a)	MCAR	-	966.3	1466.7	0.0081	-		0.0006
2(b)	-	biv. norm.	969.1	1449.6	-	0.0961		0.0003
2(c)	-	biv. gam.	915.4	1367.9	-	0.0853		0.0016
3(a)	UCAR	biv. norm.	970.3	1450.7	0.8690	0.0917		0.0003

3(b)	MCAR	biv. norm.	974.2	1458.3	0.0073	0.0266	0.0003
3(c)	UCAR	biv. gam.	962.6	1446.3	0.3008	0.1289	0.0003
3(d)	MCAR	biv. gam.	973.5	1453.8	0.0069	0.0627	0.0003

Process 2: bivariate gamma

Model	Random Effects		Model Fit		EB RE Corr		EB
	Spat.	Non-Spat.	DIC	MSPE	Spatial	Non-spatial	RR corr
True	-	biv. gam	-	-	0	0.5985	0.5985
1(a)	-	univ. gam.	493.7 (13)	38.3 (100)	-	0.4205 (0.0127)	0.4205 (0.0127)
1(b)	UCAR	-	510.0 (1)	43.6 (98)	0.3594 (0.0302)	-	0.3906 (0.0305)
2(a)	MCAR	-	499.6 (7)	41.9 (94)	0.7271 (0.0338)	-	0.7154 (0.0332)
2(b)	-	biv. norm.	485.9 (62)	41.6 (100)	-	0.6300 (0.0261)	0.6260 (0.0252)
2(c)	-	biv. gam.	482.3 (98)	36.7 (100)	-	0.6988 (0.0219)	0.6988 (0.0219)
3(a)	UCAR	biv. norm.	486.1 (59)	37.3 (100)	0.0036 (0.0035)	0.6294 (0.0265)	0.6220 (0.0248)
3(b)	MCAR	biv. norm.	492.5 (3)	37.7 (100)	0.3551 (0.0275)	0.5545 (0.0328)	0.5641 (0.0205)
3(c)	UCAR	biv. gam.	482.5 (99)	36.7 (100)	0.0024 (0.0029)	0.7016 (0.0226)	0.6956 (0.0218)
3(d)	MCAR	biv. gam.	490.3 (0)	37.5 (100)	0.3018 (0.0167)	0.7213 (0.0264)	0.6316 (0.0183)

Model	Random Effects		Model Fit		EB RE Corr		EB
	Spat.	Non-Spat.	DIC	MSPE	Spatial	Non-spatial	RR corr
1(a)	-	univ. gam.	61.7	50.7	-	0.0335	0.0335
1(b)	UCAR	-	57.9	47.6	0.1470	-	0.0506
2(a)	MCAR	-	61.9	50.8	0.5399	-	0.0160
2(b)	-	biv. norm.	63.6	52.4	-	0.0052	0.0017
2(c)	-	biv. gam.	67.7	55.0	-	0.0111	0.0111
3(a)	UCAR	biv. norm.	63.6	52.4	0.0001	0.0052	0.0015
3(b)	MCAR	biv. norm.	65.3	53.8	0.1420	0.0096	0.0022
3(c)	UCAR	biv. gam.	67.8	55.0	0.0001	0.0117	0.0174
3(d)	MCAR	biv. gam.	69.3	56.6	0.0979	0.0190	0.0022

^aTwo disease-specific univariate gamma random effects.

^bTwo disease-specific univariate conditionally autoregressive random effects.

Author Manuscript

c Two multivariate conditionally autoregressive random effects.
 d Two bivariate normal random effects.
 e Two bivariate gamma random effects.

Author Manuscript

Author Manuscript

Author Manuscript