Disease mapping of zero-excessive mesothelioma data in Flanders

Peer-reviewed author version

NEYENS, Thomas; LAWSON, Andrew; Kirby, Russell S.; Nuyts, Valerie; WATJOU, Kevin; AREGAY, Mehreteab; Carroll, Rachel; NAWROT, Tim & FAES, Christel
(2017) Disease mapping of zero-excessive mesothelioma data in Flanders. In: ANNALS OF EPIDEMIOLOGY, 27(1), p. 59-66.

DOI: 10.1016/j.annepidem.2016.10.006
Handle: http://hdl.handle.net/1942/23749

Disease Mapping of Zero-excessive Mesothelioma Data in Flanders

Thomas Neyens[1], Andrew B. Lawson[2], Russell S. Kirby[3], Valerie Nuyts[4], Kevin Watjou[1], Mehreteab Aregay[2], Rachel Carroll[2], Tim S. Nawrot[4,5], Christel Faes[1]

[1] *I-BioStat, University of Hasselt, Hasselt, Belgium*

[2] *Division of Biostatistics and Epidemiology, College of Medicine, Medical University of South Carolina, Charleston, SC, USA*

[3] *Department of Community and Family Health, College of Public Health, University of South Florida, Tampa, FL, USA*

[4] *Centre for Environment and Health, Department of Public Health and Primary Care, KU Leuven, Leuven, Belgium*

[5] *Centre for Environmental Sciences, University of Hasselt, Hasselt, Belgium*

Abstract

Purpose To investigate the distribution of mesothelioma in Flanders using Bayesian disease mapping models that account for both an excess of zeros and overdispersion.

Methods The numbers of newly diagnosed mesothelioma cases within all Flemish municipalities between 1999 and 2008 were obtained from the Belgian Cancer Registry. To deal with overdispersion, zero-inflation and geographical association, the hurdle combined model was proposed, which has three components: a Bernoulli zero-inflation mixture component to account for excess zeros, a gamma random effect to adjust for overdispersion and a normal conditional autoregressive random effect to attribute spatial association. This model was compared with other existing methods in literature.

Results The results indicate that hurdle models with a random effects term accounting for extra-variance in the Bernoulli zero-inflation component fit the data better than hurdle models that do not take overdispersion in the occurrence of zeros into account. Furthermore, traditional models that do not take into account excessive zeros but contain at least one random effects term that models extra-variance in the counts have better fits compared to their hurdle counterparts. In other words, the extra-variability, due to an excess of zeros, can be accommodated by spatially structured and/or unstructured random effects in a Poisson model such that the hurdle mixture model is not necessary.

Conclusions Models taking into account zero-inflation do not always provide better fits to data with excessive zeros than less complex models. In this study, a simple conditional autoregressive model identified a cluster in mesothelioma cases near a former asbestos processing plant (Kapelle-op-den-Bos). This observation is likely linked with historical local asbestos exposures. Future research will clarify this.

Some Keywords: Excess Zeros; Mesothelioma; Disease Mapping; Conditional Autoregressive Convolution Model; Bayesian Analysis

List of Abbreviations

Here, a list is given of the most often used abbreviations in this manuscript.

****	:	Poisson model
**G*	:	Poisson-gamma model
**N*	:	Poisson-lognormal model
*C**	:	conditional autoregressive model
CG	:	combined model
CN	:	conditional autoregressive convolution model
CAR	:	conditional autoregressive
CARCON	:	conditional autoregressive convolution
CH	:	correlated heterogeneity
CPO	:	conditional predictive ordinate
DIC	:	deviance information criterion
H***	:	hurdle Poisson model without random effect in zero component
H*G*	:	hurdle Poisson-gamma model without random effect in zero component
H*N*	:	hurdle Poisson-lognormal model without random effect in zero component
HC**	:	hurdle conditional autoregressive model without random effect in zero component
HCG*	:	hurdle combined model without random effect in zero component
HCN*	:	hurdle conditional autoregressive convolution model without random effect in zero component
H**N	:	hurdle Poisson model with random effect in zero component
H*GN	:	hurdle Poisson-gamma model with random effect in zero component
H*NN	:	hurdle Poisson-lognormal model witH random effect in zero component
HC*N	:	hurdle conditional autoregressive model with random effect in zero component
HCGN	:	hurdle combined model with random effect in zero component
HCNN	:	hurdle conditional autoregressive convolution model with random effect in zero component
M	:	marginal predictive likelihood
RR	:	relative risk
SIR	:	standardized incidence ratio
UH	:	uncorrelated heterogeneity

1 Introduction

Mesothelioma is an asbestos-related malignancy that mainly develops in the pleural cavity of the lungs but can also occur in the pericardium, peritoneum and tunica vaginalis. Increased risks for mesothelioma cancer have been reported for employees in occupations involving inhalation of asbestos dust, such as workers in asbestos mines, shipyards, railways, and others (WHO, 1998; Agudo et al., 2000; Magnani et al., 1995). However, individuals living in the proximity of factories that use asbestos might also have an increased risk (Magnani et al., 1995, 2000; Browne and Goffe, 1984). In Belgium, the yearly mean asbestos consumption per capita from 1960 to 1969 was the highest observed world-wide with 5.5 kg/head/year (Nawrot et al., 2007). Because of the high asbestos exposure in the past and a latency of 20 to 40 years for the disease to occur, mesothelioma incidence in Belgium is still high, with 173 cases in 2013 (Belgian Cancer Registry, 2015). Population cancer-registry data on mesothelioma have been used in many countries to investigate the spatial and/or temporal trends of the incidence of mesothelioma, e.g. in Brazil (Algranti et al., 2105), California (Pan et al., 2005), Canada (Krupoves et al. 2015), the Netherlands (Segura et al., 2003) and Spain (Lopez-Abante et al., 2005). In these studies, a higher risk for mesothelioma was seen in places with asbestos usage in the past. However, a geographical analysis of the incidence of mesothelioma is hindered by the large variability in the cases between areas, as well as the excessive number of areas with no mesothelioma cases. Van den Borre and Deboosere (2014) studied the spatial distribution of mesothelioma in Belgium and saw elevated standard mortality ratio's (SMR) in Sint-Niklaas, Mechelen, Dendermonde, Halle-Vilvoorde and Antwerp for men and in Mechelen and Halle-Vilvoorde for women. Asbestos consuming companies, shipbuilding companies and an international port, which were located in these towns, could explain the asbestos exposure. Other specific studies about temporal trends in Belgium have not yet been published, but a European incidence peak is predicted for 2020 (Opitz, 2014).

Many modern disease mapping techniques provide ways to extend the Poisson model for count data in order to deal with the spatial structure in the data and the occurrence of overdispersion.

Overdispersion means that the variability in the data is not equal to the mean as prescribed by the Poisson distribution, e.g. due to the exclusion of important covariates, and is often referred to as nonspatial variability. Spatial dependency in the sense that the relative risks in areas that are close in distance are more similar than in areas further apart, is referred to as spatially structured variability. Lawson (2013, chapter 5) gives an overview on the use of different forms of extra-variability in disease mapping models. The use of Bayesian models that allow for both spatially structured and nonspatial variability (Clayton and Kaldor, 1987; Besag et al., 1991) are very well known to investigate the geographical distribution of the disease burden. Another extension of the Poisson model, mostly for rare diseases, allows for the occurrence of extra zeros that can not be fully addressed by the Poisson model. The additional variability caused by an excessive number of zeros is regularly accounted for using so-called zero-inflated or zero-truncated models. Zero-inflated models were studied for univariate count data by Lambert (1992) and Greene (1994), with an extension towards the hierarchical setting studied in Min and Agresti (2005) and Lee et al. (2006). A zero-truncated model, the so-called hurdle model, was proposed by Mullahy (1986) and also extended to complex data settings, e.g. by Scheel et al. (2013). In the context of spatial count data, Agarwal et al. (2002; 2006), Gschössl and Gzado (2008) and Ugarte et al. (2004) used a zero-inflated spatial Poisson model.

Neyens et al. (2012, 2016) suggested the use of the so-called combined model for overdispersed spatial count data. The combined model literally forms a combination of the generalized linear mixed model that deals with structural aspects of the data through the inclusion of normal random subject-specific effects on one hand (Engel and Keen, 1992; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Molenberghs and Verbeke, 2005) and an overdispersion model, such as, for count data examples, the negative-binomial model (Breslow, 1984; Lawless, 1987), where the natural parameter is assumed to follow a gamma distribution on the other. Since the combined model in theory only accommodates clustering and overdispersion through two separate sets of normal and gamma random effects in a Poisson model, it is worthwhile to investigate a more general framework for data in which correlation, overdispersion and an excess of zeros can appear together. The usability

of the hurdle combined model will be investigated.

In this paper, we apply these methods to study the geographical distribution of mesothelioma in Flanders (Belgium). Interest is in the geographical risk map of the disease burden, since this may provide clues about clusters or hot spots of mesothelioma. Section 2 gives an overview of existing and new methods to deal with different forms of extra-variance. In Section 3, results of the application of these models on Flemish mesothelioma data are presented, while the use of hurdle models in disease mapping is discussed in Section 4. Concluding remarks are given in Section 5.

2 Material and Methods

2.1 Mesothelioma in Flanders

The Flemish mesothelioma data consist of yearly counts of newly diagnosed mesothelioma cases (of pleura, peritoneum and pericardium) for males and females from 1999 to 2008 in Flanders (without Brussels) and were collected by the Belgian Cancer Registry. The data include the observed and expected number of mesothelioma cases in each municipality, where the expectation is calculated according to an indirect gender-age standardization. Waller and Gotway (2004) provide in-depth explanations of this and other standardization techniques. This very rare but highly aggressive cancer that affects the membrane lining of the lungs, the pericardium, peritoneum or tunica vaginalis is typically linked to asbestos exposure. In Flanders, mainly Eternit NV, an asbestos cement manufacturing company, which was based in Kapelle-op-den-Bos, a small town close to Antwerp, has been notorious for using asbestos until 1994. These data are part of a larger study aimed at determining long-term effects of asbestos exposure and contamination, since after 1994 mesothelioma incidences in and around Kapelle-op-den-Bos have remained frequent, possibly caused by Eternit-sourced asbestos exposure. It is important to keep in mind that the observed counts used here reflect the locations where diagnosis and therefore not necessarily the exposure to asbestos occurred. Although the association between the occurrence of mesothelioma and the proximity of a patient's birthplace to asbestos exposure has been proved (Bayram et al., 2013), bias may exist as patients are often diagnosed with

the disease in different municipalities as where they came in contact with asbestos, especially when acknowledging the long incubation time of mesothelioma. Standardized incidence ratios (SIR) were calculated as the ratio of observed and expected cases (Figure 4 shows the SIR's for the years 1999 and 2008). According to these maps, risks are elevated in the central part of Flanders, namely in and around Antwerp and Kapelle-op-den-Bos. Other municipalities have occasionally elevated risks, but throughout the study period, the central part of Flanders protrudes above all as an area with high risks. There seems to have been little change throughout the years, not only in the spatial distribution of the disease, but also in terms of numbers of new diagnoses, which is probably due to the long incubation time of mesothelioma, even after banning the use of asbestos in the early nineties.

It seems relevant to investigate spatial correlation, since the presence of the disease is likely to be strongly correlated with asbestos exposure. Furthermore a sample average and standard deviation of 0.24 and 1.89 respectively and a large amount of zero counts (84$\%$) make considering a zero-inflated or hurdle combined model useful. Table 1 and Figure 4 show the summary statistics of the number of mesothelioma cases in the Flemish municipalities. Due to the disease being very rare, many zero counts occur. As zero-inflation is present, possibly alongside spatial effects, unobserved variables, or both, a model that accommodates excessive zeros, spatial correlation and unstructured overdispersion may be helpful.

2.2 The Combined Model for Spatial Data

Disease mapping models are used to link the observed counts Y_i for a spatial (lattice) location $i = 1, ..., n$ to the expected counts E_i and they mainly differ in how the smoothing of extra variation seen in Y_i in comparison to E_i is handled. Let ω_i denote the unknown relative risk (RR) for the ith area ($i = 1, \ldots, n$). The combined model as proposed by Neyens et al. (2012) for spatial lattice

data can be presented as follows:

$$\begin{aligned} Y_i &\sim \text{Poisson}(E_i\kappa_i\theta_i), \\ \kappa_i &= \exp\left(\xi_0 + \boldsymbol{x}_i'\boldsymbol{\xi} + u_i\right), \\ \theta_i &\sim \text{gamma}(\alpha, \beta), \end{aligned} \tag{1}$$

with an intercept ξ_0, known regressors $\boldsymbol{x}_i$ and their coefficients $\boldsymbol{\xi}$ and $\omega_i = \kappa_i\theta_i$. θ_i is a gamma-distributed random effect with shape and rate parameters α and β to capture uncorrelated heterogeneity (UH). Similar to common practice in the frailty context (Duchateau and Janssen, 2008), we assume $\alpha = \beta$. This standardizes the gamma random effect to mean 1. Spatially correlated heterogeneity (CH) is accommodated by u_i, an intrinsic conditional autoregressive (CAR) random effects term, such as introduced by Besag and Kooperberg (1995),

$$\begin{aligned} u_i|u_{k,i\neq k} &\sim N\left(\bar{\mu}_i, \sigma_i^2\right), \\ \bar{\mu}_i &= \frac{1}{\sum_{k=1}^N w_{ik}} \sum_{k=1}^N w_{ik}u_k, \\ \sigma_i^2 &= \frac{\sigma_u^2}{\sum_{k=1}^N w_{ik}}. \end{aligned} \tag{2}$$

Here, $w_{ik} = 1$ if areas i and k are adjacent and 0 otherwise. The CAR random effect is normally distributed with the mean and variance being weighted with the means and variances of adjacent areas. Although the weighting scheme presented above is the most common one, others can be applied too. Bivand et al. (2008) provide an in-depth overview on different methods to deal with this issue. Indeed, this model is closely related to the CAR convolution (CARCON) model, which uses a normal random effect in the linear predictor to capture UH instead of a gamma random effect. The CAR convolution model is given by

$$\begin{aligned} Y_i &\sim \text{Poisson}(E_i\omega_i), \\ \omega_i &= \exp\left(\xi_0 + \boldsymbol{x}_i'\boldsymbol{\xi} + u_i + v_i\right), \\ v_i &\sim N(0, \sigma_v^2) \end{aligned} \tag{3}$$

with u_i as in (2). Note that a convolution model has two random effects that need to be estimated, while only their sum is identifiable. Therefore, Leroux et al. (1999) proposed an alternative CAR

prior specification to model spatially correlated and uncorrelated heterogeneity with only 1 set of random effects. The Leroux CAR model will be used to investigate whether the use of the rather strong ICAR assumptions in the convolution models is justified.

The combined model (1) can be a valuable alternative to the commonly used CAR convolution model (3), since the results presented by Molenberghs et al. (2010) and Neyens et al. (2012, 2016) show that the gamma distribution is able to model extra-variance very well.

2.3 The Combined Hurdle Model for Spatial Data

While overdispersion models can deal with some extra-variance caused by a large amount of zeros, specific models for that case have been developed, with a hurdle model being one of them (Mullahy, 1986). The hurdle model provides a way of modeling count data using a two-part approach, whereby the first part is a Bernoulli model, controlling the probability π_i that the number of cancer cases in an area is either zero ($\pi_i = P(Y_i = 0)$) or is positive ($1 - \pi_i = P(Y_i > 0)$). Given the value is positive, a count distribution f_i, such as a Poisson distribution, is fitted for Y_i in the second part. Note however, that this distribution must be truncated at zero, as we know that $Y_i > 0$ in this case. The distribution of this model can be summarized as

$$p(Y_i = y_i|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_i, \pi_i) = \begin{cases} \pi_i & \text{if } y_i = 0, \\ (1-\pi_i)\frac{f_i(y_i|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_i)}{1-f_i(0|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_i)} & \text{if } y_i > 0, \end{cases}$$

where $\boldsymbol{b}_{1i}$ collects all normal random effects, such as a CH CAR term or an UH normal random effect. The zero-inflation component $\pi_i = \pi(\boldsymbol{x}'_{2i}\boldsymbol{\gamma} + \boldsymbol{z}'_{2i}\boldsymbol{b}_{2i})$ is modeled using a Bernoulli model: in the simplest case with only an intercept, but potentially containing known regressors $\boldsymbol{x}_{2i}$ and $\boldsymbol{z}_{2i}$, a vector of zero-inflation coefficients $\boldsymbol{\gamma}$ to be estimated, as well as normal random effects $\boldsymbol{b}_{2i}$. Common link functions, such as the logit or probit, can be used. The regressors in the count and zero-inflation component can either be overlapping, a subset of the regressors can be used for the zero-inflation, or entirely different regressors for the two parts can be used. Merging ideas of the combined model of Molenberghs et al. (2010) and the hurdle model (Mullahy, 1986), a two-part hurdle combined model is considered to deal with zero-inflated overdispersed clustered count data. While the first part

models only the zero state with probability π_i, the second part handles non-zero counts, which are assumed to follow a truncated-at-zero probability mass function, such as, in this case, a truncated Poisson-normal-gamma model.

2.4 Data Application and Model Comparison

It is important to remind the reader that we have multiple observations per municipality (denoted further on as repetitions of the spatial phenomenon). In other words, the models are fitted with data for municipality $i = 1, ..., 308$ and for repetition $j = 1, ..., 20$, since we have male and female data for 10 years, which results in $N = 6160$ observations. Note that this data set can be analysed by a spatio-temporal model. However, in this study, we assumed the relative risks to be constant over time and gender. This is a strong assumption, but it was made to simplify the model. This is justifiable, since mesothelioma, with its long incubation time, has to be studied over longer time periods to assess longitudinal patterns. Let Y_{ij} and E_{ij} respectively represent the observed and expected number of newly diagnosed mesothelioma cases. A model that accounts for excess zeros, overdispersion and a spatial trend seems reasonable when looking for SIR estimates that show spatial trends and many zeros (Figure 1). A full model for excessive zero counts can be presented as a model with mean $\mu^c_{ij} = \mu^c_i = \kappa_i \theta_i$:

$$\kappa_i \quad = \quad \exp\left(\xi_0 + u_i\right),$$

with the zero part probability $\pi_{ij} = \pi_i$ modeled as $\text{logit}(\pi_i) = \xi_1 + v_i$. The random effects term u_i introduces the spatial CAR structure into the count component, while v_i is an unstructured and normally distributed random effect in the excess zero-component. A CARCON model emerges when θ_i is assumed to be equal to 1 and when a normal UH random effect v_{1i} is added to the linear predictor of the counts (the normal UH term in the Bernoulli component is then denoted as v_{2i}). Note that in terms of random effects attributing certain sources of extra-variance, other parametrizations are possible. The particular choice here was primarily made in order to work with a concise set of

parameters. Lastly, note that no covariates, such as year and gender, were used in the models in this paper. However, time and gender effects were tested in additional analyses, but their effects were never significant. Priors were $\xi_0 \sim \mathsf{N}(0, \sigma^2 = 10000)$, $\xi_1 \sim \mathsf{N}(0, \sigma^2 = 10000)$. Furthermore, a gamma$(0.5, 0.0005)$ was assigned as a prior for precisions τ_{v1}, τ_{v2} and τ_u. The gamma parameters in the combined and Poisson-gamma models received priors $\alpha = \beta \sim \exp(1)$. A selection of codes is provided in Appendix 1.

The DIC (deviance information criterion) as proposed by Spiegelhalter at al. (2002) is a commonly used method to compare models. However, since hurdle models are mixture models and non-excess-zero models are not, DIC values, which indicate model complexity, can not be compared for model selection, as they will typically increase heavily in mixture models. Delorio and Robert (2002) provide a discussion about issues with the use of DIC in mixture models. An alternative approach is to use the conditional predictive ordinate (CPO; e.g. in Dey et al., 1997). The inverse of the CPO can be computed in WinBUGS for each observation, which can be inverted afterwards to obtain the CPO values. The marginal predictive likelihood (M) can then be calculated by summing the natural logarithms of the CPO values for all observations and can be used as a global goodness-of-fit statistic (larger is better). Note however, that similar to DIC-based model selection, there are no formal tests or guidelines on how large differences between M values have to be in order to indicate improved fits.

3 Results

When looking at the results in Table 2, it becomes clear that hurdle models that take into account extra-variance in the Bernoulli zero-inflation component (upper panel of Table 2) have better fits than their hurdle counterparts that assume the zero-inflation to be constant through space and time (middle panel of Table 2) in terms of the M statistic. Within both types of hurdle models (in other words, within both upper and middle panels), models using a normal UH term in the Poisson component (e.g., when applying the notations in Table 2, HCNN and H*NN in the upper panel) have

similar, but mostly slightly better fits than the models using a gamma-distributed UH random effect (e.g. HCGN and H*GN in the upper panel). Furthermore, convolution models (e.g. HCGN and HCNN in the upper panel) do not have improved fits when compared to models only using one CAR random effects term in the Poisson component (e.g. HC*N in the upper panel). Noteworthy is that the removal of both random effects in the Poisson linear predictor severely worsens fits (e.g. H**N vs. the other models in the upper panel).

In this analysis, the hurdle models have worse fits than the traditional models in the lower panel of Table 2. E.g. a clear improvement in model fit occurs when dropping the hurdle specification in the hurdle combined model from the upper panel (HCGN) to obtain the traditional combined model (*CG*). Even a simple Poisson model (**** in the lower panel) has a better fit than its hurdle counterparts (H**N and H***). The best fitting model would have to be picked from the lower panel, with the CAR model (without UH term; *C**) having the largest M value, almost equal to that of the CAR convolution model (*CN*).

In terms of parameter estimation, no striking results are seen. The estimated random effects standard deviations do not change drastically between traditional models and their hurdle counterparts (across columns). Though, an exception can be seen for the HCNN and HCN* models, where the former attributes proportionally more extra-variance to the CH term as compared to the UH term, while the opposite is seen for the HCN* model. It is also clearly visible that random effects standard deviations in the Poisson component of the convolution models are lower than in models that only include one random effects term in the Poisson component. Furthermore, the predicted number of zeros remains practically the same in all models and only decreases slightly when a simple Poisson model is used.

When investigating the relative risk map, estimated by a traditional CAR model (*C**) in Figure 3, a strong elevation in relative risks can be seen in and around the area of Kapelle-op-den-Bos. As was

suggested from the exploratory analysis, the central part of Flanders generally has increased risks, when compared to the eastern and western parts of Flanders.

3.1 Sensitivity Analysis

A sensitivity analysis was undertaken on the final model, to investigate the robustness of our results and verify assumptions on which these analyses were based. First of all, we investigated the sensitivity of the prior on the inference. Instead of the previously specified vague gamma prior, an improper uniform distribution was set as prior for the precision parameters (similar to specifications in Adin et al. (2016)). Parameter estimates for the standard deviation of the CAR random effect were very similar (resp. $\hat{\sigma_u} = 0.875$, s.e. $= 0.075$ vs. $\hat{\sigma_u} = 0.886$, s.e.$= 0.074$) and relative risk estimates were almost exactly the same for both models. Secondly, to investigate the impact of the specification of the ICAR random effect, results coming from our final CAR convolution model were compared with those coming from a model with a Leroux CAR model (Leroux et al., 1999; Appendix 2, Fig. A.1), in an analysis in which all male and female disease incidences were aggregated over the ten year period. The estimated autocorrelation parameter $\hat{\rho} = 0.840$ (s.e.$= 0.095$) was close to 1 and DIC fits were almost the same (final CAR model DIC$= 1224.09$, pD $= 108.20$ vs. Leroux CAR model DIC$= 1225.62$, pD $= 113.91$), which justifies the use of the CAR convolution model. Thirdly, our analyses assume a constant risk in time. An analysis with the disease cases summed up during the whole time period is possible as well, with information loss as a disadvantage however. We compared the final CAR model using data per year with a CAR model for the overall number of cases, which led to exactly the same results (Appendix 2, Fig. A.2). Lastly, to investigate whether there were spatially varying differences in the risk between males and females, a model was fitted for males and females separately. This was done since our joint analysis assumed the spatial trends for males and females to be the same. In these additional analyses, only slight differences were seen between relative risk estimates of males or females (estimates were not significantly different between males and females in more than 95% of the municipalities), which ratifies the joint analysis of males and females (Appendix 2, Fig. A.3 and A.4).

4 Discussion

It can be seen from the results in Section 3 that for the mesothelioma data, models including terms for overdispersion and spatial correlation are capable of capturing extra-variance that is present in the data, but which is caused by excessive zeros. If a hurdle model is used, fits improve greatly when an unstructured random effect is added to the Bernoulli-distributed zero-component. It is important to highlight that predictions of the percentage of zeros given by a very simple Poisson-gamma model (**G*) are almost equal to those given by complex hurdle models. From model comparison a model that takes into account spatial correlation through a CAR random effect was preferred over a CAR-CON model which had an almost identical M statistic, but was more (and possibly unnecessarily) complex. The CAR model provides us with relative risk estimates in Figure 3, in which a (slight) spatial effect can be seen, as the estimates are elevated in the central part of Flanders. Although the model only contains a spatial random effects term, the relative risk map does not become spatially over-smoothed, which gives a realistic image.

As it is seen in this study, the use of zero-truncated models is not always preferable, even if the data suffer from excessive zeros. The most important reason is that through the use of one or multiple random effects a lot of extra-variability can be captured, often regardless of the underlying source of variability. E.g. the gamma and normal random effects in the combined model have already proven to be very flexible in capturing additional extra-variability (Neyens et al., 2012). Therefore, it may not be necessary to use a more complex model, i.e. a zero-truncated combined model, to augment the fit. Another important part of the explanation is given by the very nature of disease mapping, in which information is limited. Especially when many zeros occur, it becomes difficult for the estimation process to assign the correct amount of extra-variability to the corresponding random effects. Furthermore, non-zero counts are generally low (Table 1 and Figure 4). A simple Poisson process with a small mean can then capture a notable amount of zero-observations, while the extra-variance induced by the excess zeros can be accounted for by the UH (or CH) random

effects. However, when the non-zero counts are generally high, a Poisson model will not as easily be able to model a large amount of zeros and the UH and CH random effects may not be sufficient to cope with the large amount of zeros. This was also seen when data were generated with 80% zeros, but with 20% non-zeros with high counts (mean $= 40$) which were concentrated in 1 Flemish province (not shown here). In that case, the hurdle models fitted better than the traditional ones.

With the use of not only one but multiple sets of random effects, one also has to reflect on the precise nature of such latent structures. As underscored by Verbeke and Molenberghs (2010), full verification of the adequacy of a random effects structure is not possible based on statistical considerations alone. Furthermore, one can question whether the data contain enough information to feed these complex random effects structures. Indeed, as seen in this case study, which was of a considerate size, complex models venture towards a border where it becomes difficult to attribute an array of different sources of variability to the corresponding elements in the model. Intuition tells us that when data sets become very large, hurdle models with multiple random effects will be good options to model data with excessive zeros, as the information contained in the data is large enough to be captured by both the zero component and random effects terms. When data sets are small or intermediate of size however, non-excess-zero models seem to be able to capture the excessive zeros well while estimation remains feasible. This leads to interesting avenues for further research, namely a simulation study to investigate which sample sizes allow for good estimation in the zero-excess models.

Van den Borre and Deboosere (2014) also analysed mesothelioma-caused deaths between 1969 and 2009 in Belgium and found links with asbestos exposure. They also concluded that districts in the Central Northern part of Belgium had elevated mortalities. This has been confirmed by our study, in which mesothelioma incidences in Flanders, the Northern part of Belgium, were investigated. The obtained results point towards elevated risks in the Central Northern part of Flanders, but add more precision, as there is an increased risk in and around Kapelle-op-den-Bos. It is however important not

to over-interpret the given results. The indirect gender-age standardization only produces rates that can be compared between areas when the rates for each area are proportional to the rates for any other area, with the same proportionality factor between them (Fleiss et al., 2003). Although there is no reason to assume that these area-specific rates are not proportional to each other in Flanders, this issue needs attention, since it may be a source of bias. This and other concerns are voiced by Ocaña-Riola (2010); e.g. The use of standardised mortality ratios (SMR) and standardised incidence ratios (SIR) has a long history in disease mapping, but they should only be used for exploratory purposes as they have a number of flaws. Commonly, the SIR is calculated as the ratio of the observed number of cases in an area and the expected number of cases according to the overall disease rate in the entire study region (see e.g. Banerjee et al. 2014 or Lawson, 2013, for a recent discussion). Similar to all population confounding, the choice of reference population and the choice of whether to age-gender standardise is one which can have impact on the appropriateness of the analysis. When calculating the SIR for a certain area, one can show that the standard or reference population is actually the population in the area itself and not the outside area as is wrongly believed by many. As a result, one has to be careful when comparing SIR's (or RR's estimated by the models) between areas, since they are based on different reference populations. While there are disadvantages of the application of the SIR, its use as an exploratory tool is generally accepted, as seen in the large amount of papers using SIR's in disease mapping.

5 Conclusions

In this paper, a modeling strategy for hierarchical count data was described where excessive zeros, correlation and overdispersion can occur together and are assembled in one single model. However, the data were modeled best by a traditional model. Indeed, extra-variability, due to an excess of zeros, could be accommodated by spatially structured random effects in a simpler Poisson model such that a hurdle mixture model for excessive zeros was not necessary to obtain better fits. From a CAR model, estimated relative risks for mesothelioma were seen in the central part of Flanders,

especially in Kapelle-op-den-Bos and in its north-western neighboring municipalities. As Kapelle-op-den-Bos is known for Eternit NV and its use of asbestos, this is a convincing result. Although it is not clear from the relative risk map in Figure 3, there are other areas in Flanders too that possibly have historically suffered from increased asbestos exposure. In future research, it will be interesting to collect information about all exposure sources and to subsequently use this material in new analyses. Furthermore, it is important to note that mesothelioma has very long incubation times (Kuproves et al., 2015). Therefore, it becomes less straightforward to link the location of asbestos exposure with the municipalities where individuals lived when they were diagnosed with the disease. Next to that, it seems interesting to focus on spatio-temporal trends in future studies. This aspect was beyond the scope of this study, as the main intention now was to investigate zero-inflated data. When keeping in mind that the disease develops very slowly, the 10 years time frame that was featured in this research might not be enough to capture existing longitudinal patterns. Therefore a study that brings together data from several decades will be needed to investigate spatio-temporal phenomena. Lastly, other correlation structures, such as the ones between male and female counts, were not investigated here due to the complexity of the models at hand. They should therefore also be taken into account in future studies.

6 Acknowledgements

Support from the National Institutes of Health is acknowledged [award number R01CA172805]. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged.

7 References

Adin, A., Martínez-Beneito, M. A., Botella-Rocamora, P., Goicoa, T., and Ugarte, M. D. (2016) Smoothing and high risk areas detection in space-time disease mapping: a comparison of P-splines, autoregressive, and moving average models. *Stochastic Environmental Research and*

Risk Assessment, doi: 10.1007/s00477-016-1269-8.

Agarwal, D.K. (2006) Two-fold spatial zero-inflated models for analysing isopod settlement patterns. In: Upadhyay, S.K., Singh, U., Dey, D.K., ed. *Bayesian Statistics and its Applications*. New Delhi: Anamaya Publishers.

Agarwal, D.K., Gelfand, A.Z., and Citron-Pousty, Z. (2002) Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, **9**, 341–355.

Agudo, A., Gonzalez, C.A., Bleda, M.J., Ramirez, J., Hernandez, S., Lopez, F., Calleja, A., Panades, A., Turuquet, D., Escolar, A., Beltran, M., and Gonzalez-Moya, J.E. (2000) Occupation and riks of malignant pleural mesothelioma: a case-control study in Spain. *American Journal of Industrial Medicine*, **37**, 159–168.

Algranti, E., Satio, C.A., Carneiro, A.P.S., Moreira, B., Mendonca, E.M.C., and Bussacos, M.A. (2015) The next mesothelioma wave: Mortality trends and forecast to 2030 in Brazil. *Cancer Epidemiology*, **39**, 687–692.

Bayram, M., Dongel, I., Bakan, N.D., Yalçin, H., Cevit, R., Dumortier, P., and Nemery, B. (2013) High risk of malignant mesothelioma and pleural plaques in subjects born close to ophiolites. *Chest*, **143**, 164-171.

Belgian Cancer Registry (2015) *Cancer burden in Belgium 2004-2013*. Brussels. `http://www.belgiancancerregistry.be/media/docs/publications/BCR_publicatieCancerBurden2016_web160616.pdf`

Besag J., York, J., and Mollie, A. (1991) Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

Bivand, R.S., Pebesma, E.J., and Gómez-Rubio, V. (2008) *Applied Spatial Data Analysis with R*. New York: Springer.

Breslow, N. E. (1984) Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**, 38–44.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Browne, K. and Goffe, T. (1984) Mesothelioma due to domestic exposure to asbestos. *BMJ (Clin Res Ed)*, **83**, 104–111.

Clayton, D. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

Delorio, M. and Robert., C.P. (2002) Discussion of Spiegelhalter et al. *Journal of the Royal Statistical Society, Series B*, **64**, 629–630.

Dey, D.K., Chen, M.-H., and Chang, H. (1997) Bayesian approach for nonlinear random effects models *Biometrics*, **53**, 1239–1252.

Duchateau, L. and Janssen, P. (2008) *The Frailty Model*. New York: Springer.

Engel, B. and Keen, A. (1992) A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, **48**, 1–22.

Fleiss, J.L., Levin, B., and Paik, M.C. (2003). The standardization of rates, in Fleiss, J.L., Levin, B., and Paik, M.C. (eds). *Statistical methods for rates and proportions (3rd Edition)*, 627-647. New Jersey: Wiley & Sons.

Greene, W. (1994) Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper EC- 94-10, Department of Economics, New York University. *Working Paper*, 9–10.

Gschössl, S. and Czado, C. (2008) Modeling count data with overdispersion and spatial effects. *Statistical Papers*, **49**, 531–552.

Krupoves, A., Camus, M., and De Guire, L. (2015)Incidence of malignant mesothelioma of the pleura in Quebec and Canada from 1984 to 2007, and projections from 2008 to 2032. *American Journal of Industrial Medicine*, **58**, 473–482.

Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Lawless, J. (1987) Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, **15**, 209–225.

Lawson, A. B. (2013) *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. Second Edition.* Boca Rotan: Chapman & Hall.

Lee, Y., Nelder, J. and Pawitan, Y. (2006) *Generalized Linear Models With Random Effect*. London: Chapman and Hall, p. 178.

Leroux, B., Lei, X., and Breslow, N. (1999) *Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence*, chapter Statistical Models in Epidemiology, the Environment and Clinical Trials, Halloran, M. and Berry, D. (eds), 135–178. New York: Springer-Verlag.

Lopez-Abente, G., Hernandez-Barrera, V., Aragones, N., and Perez-Gomez, B. (2005) Municipal pleural cancer mortality in Spain. *Occupational and Environmental Medicine*, **62**, 195–199.

Magnani, C., Agudo, A., Gonzalez, C.A., Andrion, A., Calleja, A., Chellini, E., Dalmasso, P., Escolar, A., Hernandez, S., Ivaldi, C., Mirabelli, D., Ramirez, J., Turuquet, D., Usel, M, and Terracini, B. (2000) Multicentric study on malignant pleural mesothelioma and non-occupational exposure to asbestos. *Br. J Cancer*, **83**, 104–111.

Magnani, C., Terracini, B., Ivaldi, C., Botta, M., Mancini, A., and Andrion, A. (1995) Pleural malignant mesothelioma and non-occupational exposure to asbestos in Casala Monferrato, Italy, **52**, 362–367.

Min, Y. and Agresti, A. (2005) Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, **5**, 1–19.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data.* New York: Springer.

Molenberghs, G., Verbeke, G., Demétrio, G., and Vieira, A. (2010) A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical science*, **25**, 325–347.

Mullahy, J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341–365.

Nawrot, T.S., Van Kersschaever, G., Van Eycken, E., and Nemery, B. (2007) Belgium: historical champion in asbestos consumption. *The Lancet*, **369**, 1692.

Neyens, T., Faes, C., and Molenberghs, G. (2012) A generalized Poisson-gamma model for spatially overdispersed data. *Spatial and Spatio-temporal Epidemiology*, **3**, 185–194.

Neyens, T., Lawson, A. B., Kirby, R. S., and Faes, C. (2016) The bivariate combined model for spatial data analysis. *Statistics in Medicine*, **35**.

Ocaña-Riola, R. (2010) Common errors in disease mapping. *Geospatial Health*, **4**, 139-154.

Opitz, I. (2014) Management of malignant pleural mesothelioma - The European experience. *Journal of Thoracic Disease*, **6 (suppl 2)**, S238-252.

Pan, X.-l., Day, H.W., Wang, W., Becket, L.A., and Schenker, M.B. (2004) Residential proximity to naturally occurring asbestos and mesothelioma risk in California. *American Journal of Respiratory and Critical Care Medicine*, **172**, 1019–1025.

Scheel, I., Ferkingstad, E., Frigessi, A., Haug, O., Hinnerichsen, M., and Meze-Hausken, E. (2013) A Bayesian hierarchical model with spatial variable selection: the effect of weather on insurance claims. *Journal of the Royal Statistical Society, Series C*, **62**, 85–100.

Segura, O., Burdorf, A., and Looman, C. (2003) Update of predictions of mortality from pleural mesothelioma in The Netherlands. *Occupational and Environmental Medicine*, **60**, 50–55.

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002)Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, **64**, 583–639.

Ugarte, M.D., Ibáñez, B., and Militino, A.F. (2004) Testing for Poisson Zero Inflation in Disease mapping. *Biometrical Journal*, **46**, 526–539.

Van den Borre, L. and Deboosere, P. (2014) Asbestos in Belgium: an underestimated health risk. The evolution of mesothelioma mortality rates (1969-2009). *International Journal of Occupational and Environmental Health*, **20**, 134–140.

Verbeke, G. and Molenberghs, G. (2010) Arbitrariness of models for augmented and coarse data, with emphasis on incomplete-data and random-effects models. *Statistical Modelling*, **10**, 391–419.

Waller, L.A. and Gotway, C.A. (2004) *Applied Spatial Statistics for Public Health Data*. New Jersey: Wiley.

Wolfinger, R. and O'Connell, M. (1993) Generalized linear mixed models. *Journal of Statistical Computation and Simulation*, **48**, 233–243.

World Health Organisation. (1998) *International programme on chemical safety. Environmental health criteria 203, Chrysotile asbestos*. Geneva.

Appendix 1: Selection of WinBUGS model codes

7.1 Hurdle combined model with UH random effect in zero component (HCGN)

```
model{
for (i in 1 :N) {

# Zeros trick:
ze[i]<-0
ze[i]~dpois(phi[i])
phi[i]<- -ll[i]+10000
zero[i] <- equals(Y[i], 0)

# Count component:
mu[i]<-E[i]*kappa[i]*theta[ID[i]]
log(kappa[i]) <- xi0 + u[ID[i]]

# Zero component:
logit(p0[i]) <- xi1+ v[ID[i]]
pzero[i]<-p0[i]

#Likelihood:
ll[i] <- zero[i]*log(p0[i])+(1-zero[i])*(log(1-p0[i])-log(1-exp(-mu[i]))
-mu[i] - loggam(Y[i] + 1) +Y[i]*log(mu[i]))

#M-Statistic:

CPinv[i]<-exp(-ll[i])
}

#Predicted zeros:

predzero<-mean(pzero[])

# UH priors:
for (l in 1:n) {
theta[l]  ~ dgamma(a,a)
v[l]  ~ dnorm(0.0, tau.v)}

# CAR prior distribution for random effects:
u[1:n] ~ car.normal(adj[], weights[], num[], tau.u)
for(k in 1:sumNumNeigh) {
weights[k] <- 1}

# Other priors:
xi0 ~dnorm(0, 0.0001)
xi1 ~dnorm(0, 0.0001)
tau.u ~dgamma(0.5, 0.0005)
sig.u<-sqrt(1/tau.u)
tau.v ~dgamma(0.5, 0.0005)
sig.v<-sqrt(1/tau.v)
a ~ dexp(1)
sd_theta<-1/sqrt(a)
}
```

7.2 Combined model (*CG*)

```
model{
for (i in 1 :N) {

# Zeros trick:
ze[i]<-0
```

```
ze[i]~dpois(phi[i])
phi[i]<- -ll[i]+10000
zero[i] <- equals(Y[i], 0)

# Combined model:
mu[i]<-E[i]*kappa[i]*theta[ID[i]]
log(kappa[i]) <- xi0 + u[ID[i]]
logomega[i] <- xi0 + u[ID[i]] + log(theta[ID[i]])
omega[i]<-exp(logomega[i])
pzero[i]<-exp(-mu[i])

# Likelihood:
ll[i] <-Y[i]*log(mu[i])-mu[i]-logfact(Y[i])

#M-Statistic:

CPinv[i]<-exp(-ll[i])
}

#Predicted zeros:

predzero<-mean(pzero[])

# UH priors:

for (l in 1:n) {
theta[l]  ~ dgamma(a,a)}

# CAR prior distribution for random effects:
u[1:n] ~ car.normal(adj[], weights[], num[], tau.u)
for(k in 1:sumNumNeigh) {
weights[k] <- 1}

# Other priors:
xi0 ~dnorm(0, 0.0001)
tau.u ~dgamma(0.5, 0.0005)
sig.u<-sqrt(1/tau.u)
a ~ dexp(1)
sd_theta<-sqrt(1/a)
}
```

Appendix 2: Sensitivity analysis results

Place Figures A.1 to A.4 here.

Tables and Figures

Table 1: Yearly (male and female) mesothelioma cases in the 308 municipalities of Flanders from 1999 to 2008: summary statistics.

Statistics	**Mesothelioma**
mean	0.24
sd	0.89
median	0
min	0
max	23

Table 2: Mesothelioma study. Parameter estimates and standard errors are given for the different fitted models. Each model is indicated by 4 symbols. The first symbol indicates whether the model is a hurdle model (H) or a traditional model that does not take into account excess zeros (*). The second symbol indicates whether a CAR random effects term (C) or no spatial term (*) is used in the Poisson component. The third symbol indicates whether a gamma (G), a normal (N) or no (*) unstructured random effect is used in the Poisson component. The fourth symbol indicates whether a normal (N) or no (*) random effects term is used in the Bernoulli component of the hurdle models. If a traditional (non-hurdle) model is used, the fourth symbol also will be a *.

		HCGN	HCNN	HC*N	H*GN	H*NN	H**N
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Poisson: intercept	ξ_0	-0.056(0.170)	-0.297(0.096)	-0.315(0.103)	0.152(0.099)	-0.196(0.117)	0.276(0.041)
Poisson: std. dev. CAR	σ_u	0.619(0.357)	1.087(0.158)	1.206(0.156)	-	-	-
Poisson: std. dev. gam. UH	$\sqrt{1/\alpha}$	0.690(0.199)	-	-	0.972(0.100)	-	-
Poisson: std. dev. norm. UH	σ_{v1}	-	0.227(0.120)	-	-	0.824(0.091)	-
Inflation: intercept	ξ_1	1.937(0.070)	1.937(0.071)	1.937(0.069)	1.942(0.070)	1.940(0.072)	1.943(0.074)
Inflation: std. dev. UH	σ_{v2}	0.961(0.063)	0.961(0.064)	0.961(0.063)	0.962(0.063)	0.963(0.065)	0.965(0.065)
Predicted prob. zeros		0.840	0.840	0.840	0.840	0.840	0.840
M		-3133.14	-3127.84	-3128.68	-3140.51	-3135.41	-3253.02
		HCG*	HCN*	HC**	H*G*	H*N*	H***
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Poisson: intercept	ξ_0	-0.180(0.113)	-0.190(0.110)	0.310(0.100)	0.151(0.0943)	-0.229(0.124)	0.176(0.042)
Poisson: std. dev. CAR	σ_u	0.914(0.126)	0.390(0.122)	1.210(0.148)	-	-	-
Poisson: std. dev. gam. UH	$\sqrt{1/\alpha}$	0.564(0.101)	-	-	0.966(0.104)	-	-
Poisson: std. dev. norm. UH	σ_{v1}	-	0.669(0.124)	-	-	0.846(0.096)	-
Inflation: intercept	ξ_1	1.661(0.035)	1.660(0.034)	1.660(0.035)	1.661(0.034)	1.660(0.035)	1.660(0.034)
Predicted prob. zeros		0.840	0.840	0.840	0.840	0.840	0.840
M		-3365.501	-3367.34	-3365.81	-3382.23	-3371.82	-3489.76
		CG	*CN*	*C**	**G*	**N*	****
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Poisson: intercept	ξ_0	-0.210(0.043)	-0.254(0.037)	-0.254(0.035)	-0.046(0.048)	-0.237(0.051)	0.004(0.026)
Poisson: std. dev. CAR	σ_u	0.693(0.087)	0.870(0.073)	0.875(0.075)	-	-	-
Poisson: std. dev. gam. UH	$\sqrt{1/\alpha}$	0.353(0.041)	-	-	0.633(0.042)	-	-
Poisson: std. dev. norm. UH	σ_{v1}	-	0.036(0.019)	-	-	0.584(0.041)	-
Predicted prob. zeros		0.840	0.840	0.840	0.838	0.839	0.823
M		-2796.93	-2789.84	-2789.75	-2831.26	-2812.52	-3034.70

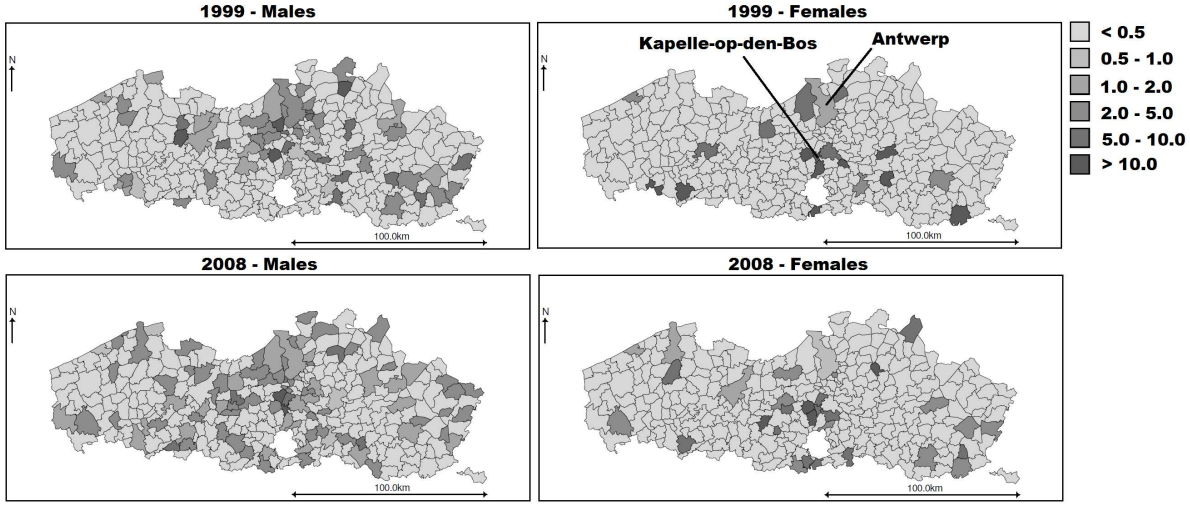

Figure 1: Standardized incidence ratio (SIR) maps of newly diagnosed male mesothelioma cases in the 308 municipalities of Flanders for males and females in 1999 and 2008.

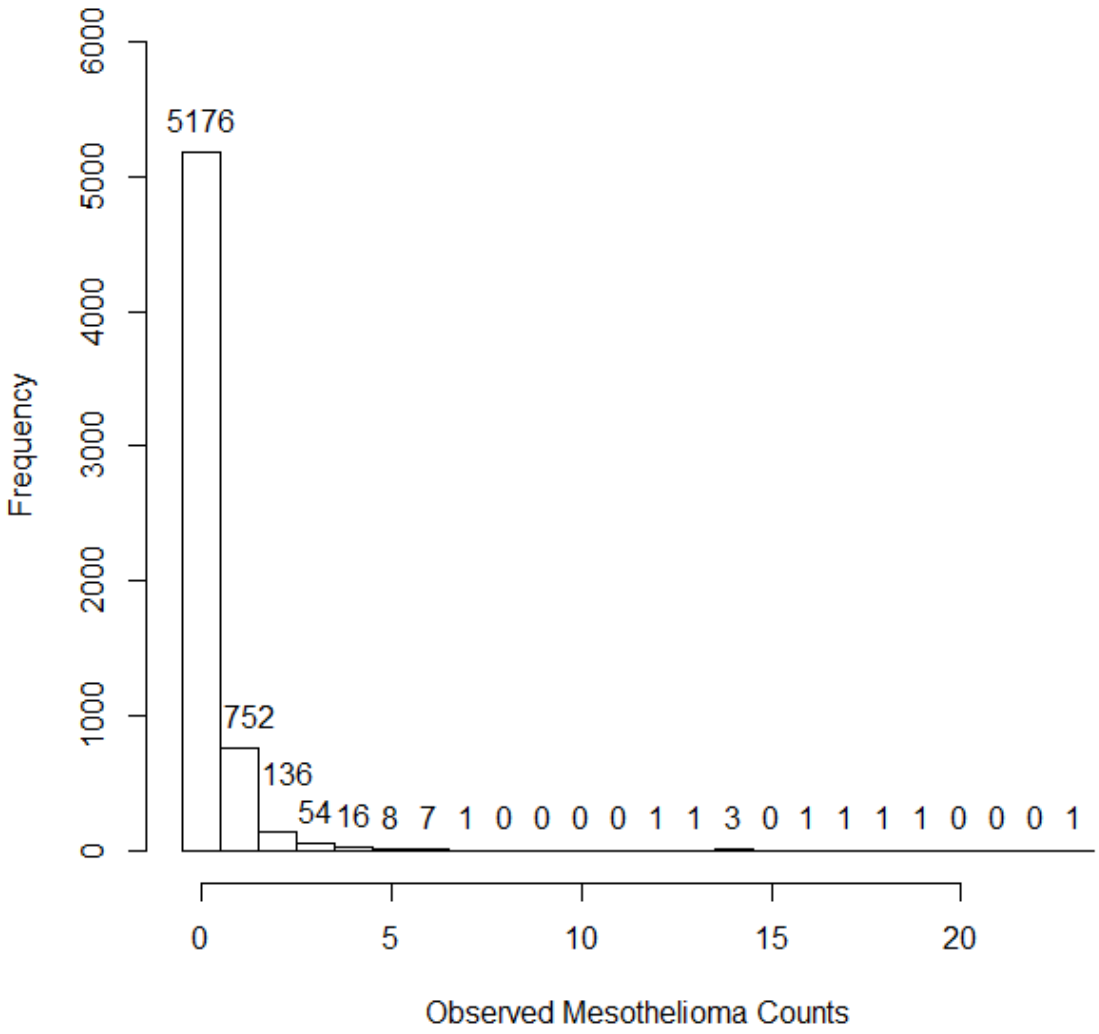

Figure 2: Histogram of observed mesothelioma cases.

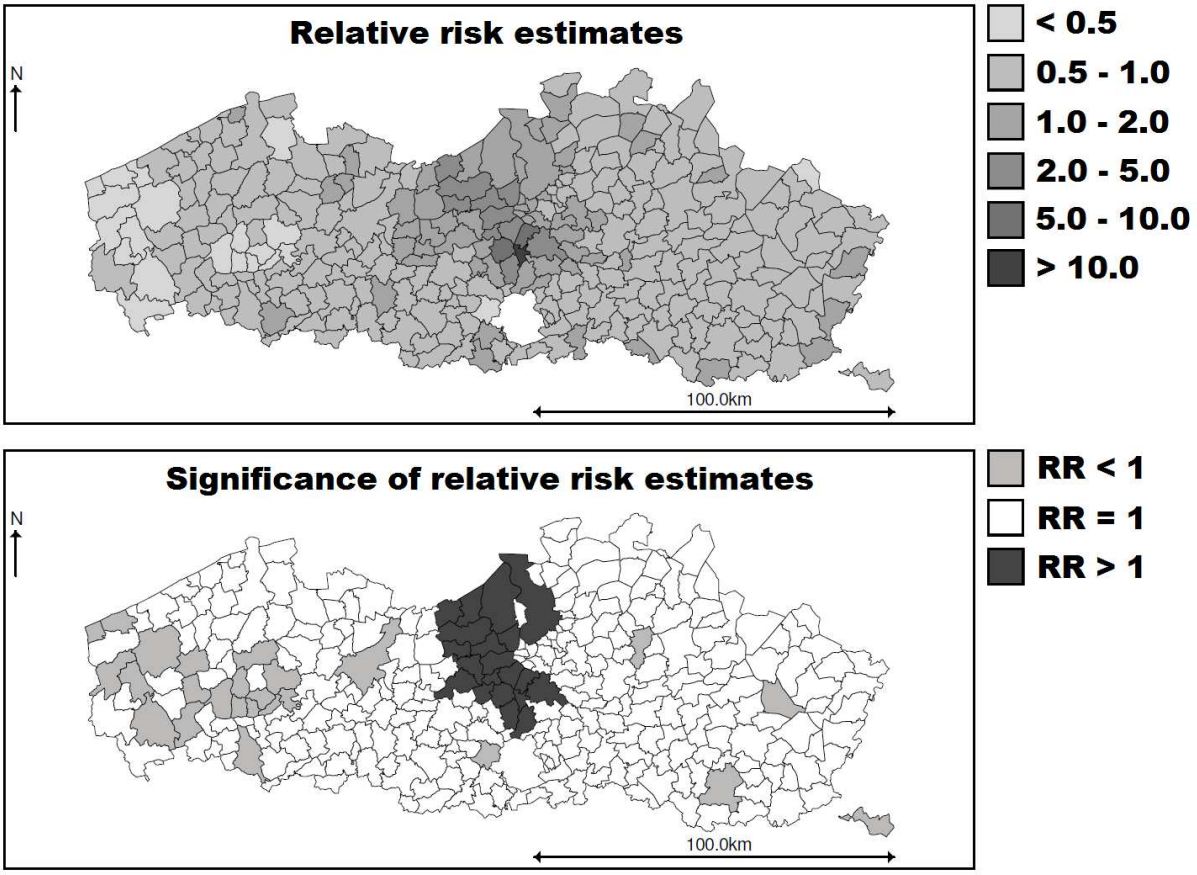

Figure 3: Relative risk estimates, given by the CAR model (*C**; upper map) and indications if locations have relative risks that differ from 1 (5% significance level), with the additional indication if the estimates are lower or larger than 1 (lower map).

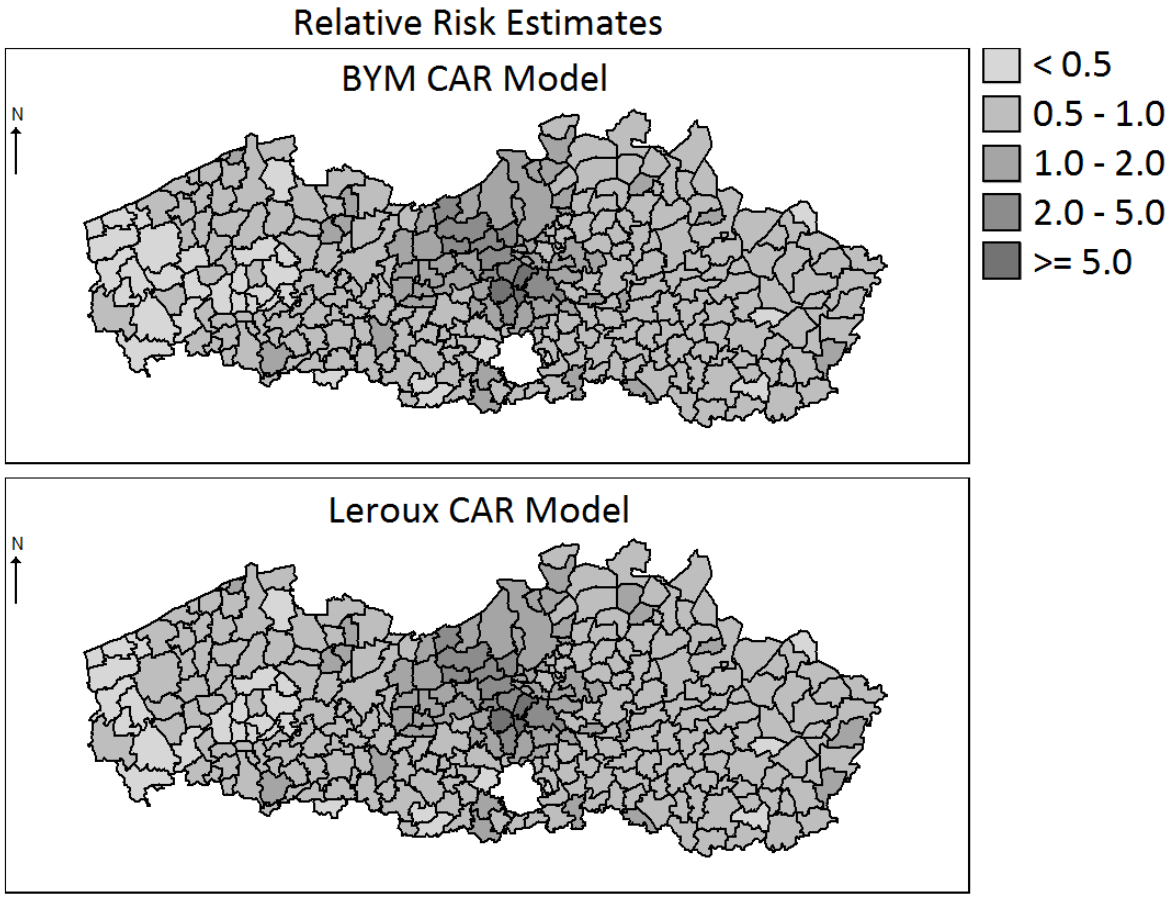

Figure A 1: RR estimates for the Leroux model vs. the final CAR model.

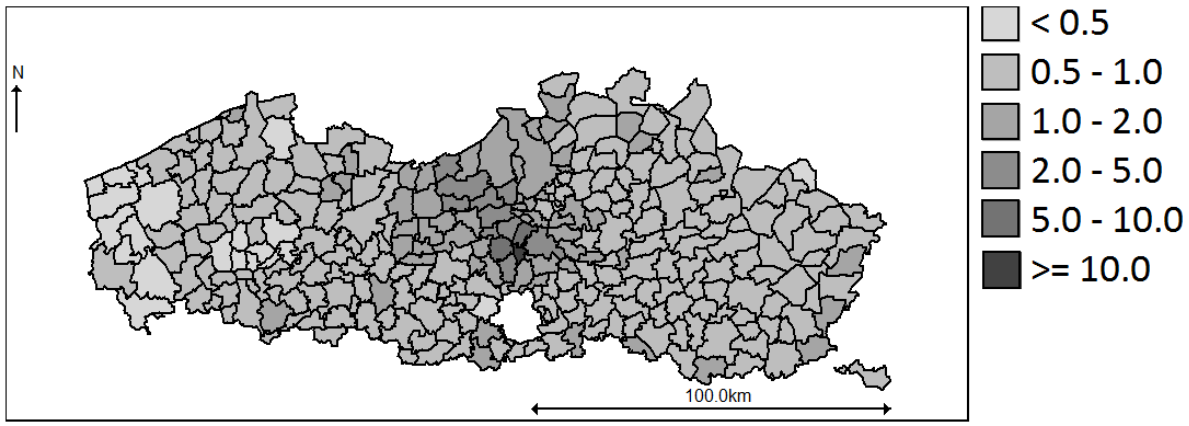

Figure A 2: RR estimates for the final CAR model, based on the aggregated data for males and females for the whole study period.

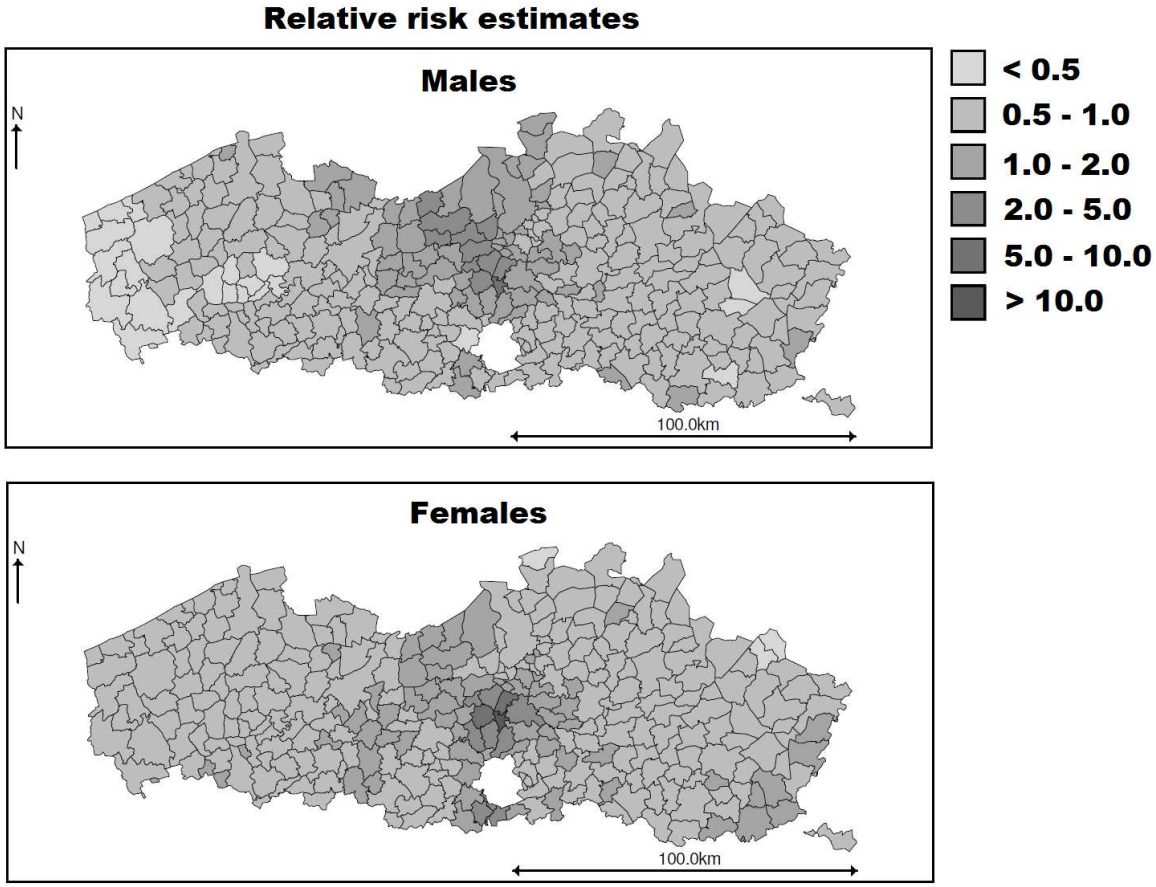

Figure A 3: RR estimates for the final CAR model, based on male and female data sets separately.

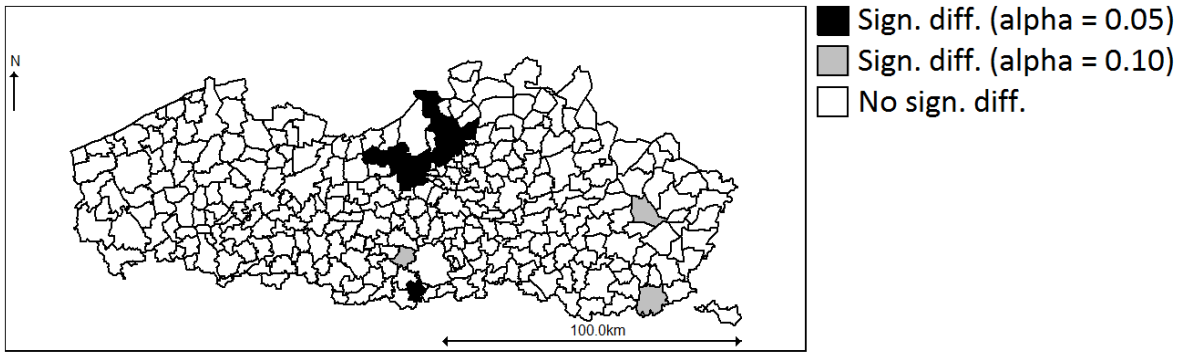

Figure A 4: Significance of differences between male and female relative risks when modeling both genders separately.