Making the most of spatial information in health: a tutorial in Bayesian disease mapping for areal data

Su Yun Kang,[1,2] Susanna M. Cramb,[1,3] Nicole M. White,[1,2] Stephen J. Ball,[4]
Kerrie L. Mengersen[1,2]

[1]Mathematical Sciences School, Queensland University of Technology, Brisbane; [2]Cooperative Research Centre Programme for Spatial Information, Melbourne; [3]Viertel Centre for Research in Cancer Control, Cancer Council Queensland, Brisbane; [4]School of Nursing, Midwifery and Paramedicine, Faculty of Health Sciences, Curtin University, Perth, Australia

Abstract

Disease maps are effective tools for explaining and predicting patterns of disease outcomes across geographical space, identifying areas of potentially elevated risk, and formulating and validating aetiological hypotheses for a disease. Bayesian models have become a standard approach to disease mapping in recent decades. This article aims to provide a basic understanding of the key concepts involved in Bayesian disease mapping methods for areal data. It is anticipated that this will help in interpretation of published maps, and provide a useful starting point for anyone interested in running disease mapping methods for areal data. The article provides detailed motivation and descriptions on disease mapping methods by explaining the concepts, defining the technical terms, and illustrating the utility of disease mapping for epidemiological research by demonstrating various ways of visualising model outputs using a case study. The target audience includes spatial scientists in health and other fields, policy or decision makers, health geographers, spatial analysts, public health professionals, and epidemiologists.

Correspondence: Su Yun Kang, Mathematical Sciences School, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia.
Tel: +61.7.31382603.
E-mail: suyun602@gmail.com

Key words: Areal data; Bayesian mapping; Disease mapping; Spatial information; Visualisation.

Acknowledgements: the work was supported by the Cooperative Research Centre for Spatial Information, whose activities are funded by the Australian Commonwealth's Cooperative Research Centres Programme. The authors are grateful to the Editor and two anonymous referees for their very useful comments that improved the quality of the paper.

Received for publication: 12 November 2015.
Revision received: 22 January 2016.
Accepted for publication: 5 February 2016.

Geospatial Health 2016; 11:428
doi:10.4081/gh.2016.428

Introduction

Disease mapping is a flourishing field due to the growing amount of routinely collected health information worldwide (Rytkönen, 2004). Advances in geographic information systems have greatly aided the analytical manipulation and visual representation of spatial data (Burrough and McDonnell, 1998). Spatial information in health is especially useful for informing the locations of disease occurrences and the onus is on making the best possible use of this information.

Some excellent introductory guides for disease mapping are available in the literature. Nonetheless, many of these are either not intended for non-statistical audiences, or lack specific details. For instance, Elliot *et al.* (2000) present a comprehensive review of the recent developments in spatial epidemiology but the statistical methods require a level of background knowledge, which may not be suitable for beginners. Marshall (1991) covers a broad range of methods for the analysis of the geographical distribution of disease, rather than upskill the reader in using particular methods. Lawson and Williams (2001) provide a broad overview of the issues concerning disease mapping but is short on specifics (English, 2001). Banerjee *et al.* (2014) present a fully model-based approach to all types of spatial data, including point level, areal, and point pattern data. Cramb *et al.* (2011b) offer insight into the decisions made in generating a health atlas, but is not intended as an entry-level article for a non-statistical audience. This article fills the niche by providing motivation, definition and description at a general level, and illustrating these ideas via a substantive case study.

Although disease mapping has been undertaken in various forms for over 100 years, the opportunity now exists to use model-based maps that acknowledge uncertainty in inputs and outputs (López-Abente *et al*., 2014; Catelan and Biggeri, 2010), take account of the spatial nature of the data to *borrow strength* from neighbouring areas in order to improve small area estimates, and can provide probability statements (Goovaerts, 2006b). In this article, we describe Bayesian disease mapping for areal data (Lawson, 2001, 2009) as an approach that addresses these issues. We focus on a running example of mapping cancer, although the methods are applicable to other diseases.

The primary purpose of this article is to provide a basic understanding of the key concepts involved in Bayesian statistical models for disease mapping of areal data. We commence with a discussion of why disease model-based mapping methods are required. Background on Bayesian methods typically used for disease mapping is then provided, and then some of the cartographic outputs commonly used are discussed, including methods for indicating statistical uncertainty in relative risk (Appendix Part D) of disease.

Case study: cancer in Australia

Cancer is now the world's and Australia's biggest killer (IARC, 2014). The number of cases diagnosed continues to increase worldwide due to population growth and aging, with the increasing prevalence (Appendix Part D) of physical inactivity, poor diet and reproductive changes (such as later parity) also contributing (Torre *et al.,* 2015). In Australia, cancer accounts for almost one-fifth (19%) of the total disease burden (AIHW, 2014).

Disparities in cancer outcomes across broad socioeconomic status and urban/rural categories have been reported internationally (Wilkinson and Cameron, 2004; Woods *et al.,* 2006; Ernst *et al*., 2010). Within Australia, there are disparities in cancer outcomes with respect to geographic remoteness and socioeconomic status (AIHW, 2014). Cancers such as cervical and lung had higher incidence (Appendix Part D) and mortality as remoteness or area-level disadvantage increased. Furthermore, the five-year relative survival from all cancers combined decreased with greater remoteness and greater socioeconomic disadvantage.

Understanding disparities in these broad areas, while useful, is unlikely to accurately reflect the heterogeneity in outcomes at the local level. Efforts to monitor and reduce cancer disparities can benefit greatly from quantifying variation across population groups and pertinent, small geographical areas. An understanding of the geographic patterns of cancer enables health decision-making by health service planners, clinicians, epidemiologists and industry groups to be more accurate and effective, for example by targeting policy development and resource allocation at areas of greater need (Mason *et al*., 1975; Kulldorff *et al*., 2006).

Cramb *et al*. (2011a) produced the first Atlas of Cancer in Queensland to describe geographical variation in cancer incidence and survival across small areas in Queensland, using routinely-collected health information from the Queensland Cancer Registry. For the first time, Bayesian model-based cancer incidence and survival maps for Queensland were systematically presented at a comprehensive level. The Atlas significantly contributed to the understanding of geographical variation of cancer incidence and survival across Queensland, and subsequently influenced government policy decisions.

Materials and Methods

Disease maps are a visual representation of disease outcomes. The use of disease maps to aid decision making in epidemiological and medical research is well recognised (Koch, 2011). Disease maps are effective tools for explaining and predicting patterns of disease outcomes across geographical space, identifying areas of potentially elevated risk, and formulating and validating aetiological hypotheses for a disease (Shen and Louis, 2000). They are able to uncover local-level inequalities frequently masked by health estimates from large areas such as states, regions or cities (Borrell *et al.*, 2010), enabling the development of disease reduction and prevention programmes targeting high-risk populations, see for instance, Mason *et al*. (1975) and Kulldorff *et al*. (2006) who have used cancer maps to depict the geographic patterns of cancer outcomes.

Disease mapping encompasses small area studies that use data aggregated over small areas and take into account local spatial correlation, see for example, Clayton and Kaldor (1987); Cressie and Chan (1989); Besag *et al.* (1991) and Bernadinelli *et al.* (1997). Data sparseness is common in small area analyses, especially when working with less common diseases. A small number of observed and expected disease occurrences leads to unstable risk estimates (Ancelet *et al*., 2012).

The problem of potentially unstable risk estimates for sparse spatial data needs to be mitigated to obtain reliable estimates. In practice, this is achieved by implementing spatial smoothing techniques. Spatial smoothing effectively *borrows strength* across small areas, so that the disease rate estimated for an area with a small population denominator would be weighted towards the estimated disease rate of neighbouring areas that have larger denominators. The estimates obtained by smoothing information from neighbouring areas are more reliable and robust due to the increased precision in the risk estimates in areas with few observations (Ancelet *et al*., 2012). In the context of disease mapping for small areas, the implementation of spatial smoothing is commonly achieved via the incorporation of a conditional autoregressive prior distribution for the spatial effects (Lee, 2011).

A disease-mapping model is essentially a regression model that links a disease outcome to a set of risk factors. An important concept in disease mapping models, which is common to many other regression models, is the use of random effects (Appendix Part D). In this context, random effects provide a way of estimating variation in disease risk between areas that is not otherwise captured by known risk factors (*e.g.* age, sex, socioeconomic status, *etc*.).

Why Bayesian?

Bayesian statistics takes its name from the English clergyman Thomas Bayes (1702-1761), although the key concepts were also contemporaneously established by Laplace and embedded in the general view of *inverse probability* at that time (Bernardo and Smith, 2009). It is an approach to data analysis that focuses on relating observed and unknown quantities using conditional probabilities, which are measures of the probability of an event given that another event has occurred.

In a Bayesian model (Appendix Part E, Box 1), an unknown parameter (Appendix Part D) is represented using a distribution rather than a single point estimate (Johnson, 2004). The model parameters have distributions and are probabilistic [*e.g.* parameters representing coefficients associated with covariates in a regression model might be given a Normal distribution (Appendix Part E, Box 2)]. These distributions are known as prior distributions. These prior distributions can be considered as representing the uncertainty about the parameter before the data are seen. The parameters in the prior distributions (*e.g.* the mean and variance of the prior on a regression coefficient) can also have distributions, which are known as hyperprior distributions. Again, these distributions also represent uncertainty about our knowledge of these values. The combination of the prior information and the data results in a posterior distribution. The posterior distribution can be thought of as a probability distribution on the values of an unknown parameter that combines prior knowledge about the parameter and the observed data. The Bayesian model thus consists of parameters related to one another in the form of a hierarchy. The complex nature of spatial data can be captured using this hierarchical structure (Appendix Part D)

(Shen and Louis, 2000; Best *et al*., 2005).

Random effects are generally included in these models. Typically, a random effect is specified as being Normally distributed, whereby a few areas are allowed to have a disease incidence much lower than expected based on these risk factors, a few areas much higher, but most are close to expected (following a bell curve). For spatial data, we assume that sites closer to each other are more similar, so we can use information from neighbouring sites to obtain better estimates of disease risk. Hence, when we fit a spatially correlated random effect, the variation at a particular site is Normally distributed relative to the mean of its neighbours. These random effects thus relate disease risk estimates to neighbouring estimates, producing a *smoothing* effect across the area of interest.

There are many reasons why the Bayesian approach is a useful framework for disease mapping. Firstly, Bayesian smoothing methods produce robust and reliable estimation of health outcomes of interest in a small area, even when based on small sample sizes (Ancelet *et al*., 2012). Within these small areas, the sample sizes are sometimes too small to yield estimates with adequate precision and reliability. Bayesian smoothing techniques improve the estimation by using information from neighbouring areas.

Secondly, the use of prior distributions (usually based on existing knowledge or expert opinion) in disease mapping models helps strengthen inferences (Appendix Part D) about the true value of the parameter and ensures that all relevant information is included (Gurrin *et al*., 2000). These can be *uninformative* (*e.g.* set to be normally distributed with a mean of zero and a very large variance) or *informative* if there is other information about the effect of this risk factor (given the other risk factors in the model). Thirdly, the Bayesian approach allows for quantification of the uncertainty related to the health estimates from the posterior distributions (Ghosh *et al*., 1999; Wakefield, 2007). Spatial uncertainties added to the resulting risk maps depict local details of the spatial variation of the risk and provide valuable information for policy makers to make decisions about thresholds and public health (ApSimon *et al*., 2002; Johnson, 2004; Goovaerts, 2006b).

Lastly, direct probabilistic statements can be made about the underlying and unobserved parameters of interest using their posterior probability distributions. In disease mapping, it might be of interest to make probability statements about areas of high risk for a disease. For instance, computing and mapping probabilities that the risk in an area exceeds certain thresholds can be done using the posterior probability distributions (Green and Richardson, 2002). This probability of exceedance can then be used to decide whether an area should be classified as having excess risk of a disease (Richardson *et al*., 2004). It is straightforward to make these kinds of statements in a Bayesian context, since they are directly obtained from the corresponding posterior distribution.

Data

Often health data are only available with location data supplied as a small area (known as areal data), rather than a street address geocoded to a latitude/longitude point. Determining the most appropriate region size to use involves several considerations (Appendix Part E, Box 3). This article focuses on the application of disease mapping methods for spatial data aggregated over small areas and omits the discussion of other forms of spatial data such as geostatistical and point patterns data. As an alternative, health outcome data may also be analysed at the individual level, while incorporating spatial information at any geographical scale such as a point or an area.

The data described in the Atlas (Cramb *et al*., 2011a) focused on Queensland cancer data aggregated to the statistical local area (SLA) level, which was the smallest area with annual population data available. However, consistent with most administrative regions, the areas are of varying sizes, and larger areas tend to dominate the map. An alternative approach is to aggregate disease data with continuous coordinate information to regular grid cells (Li *et al*., 2012a, 2012b; Kang *et al*., 2014). Such an approach allows modelling of disease data at a fine spatial scale, independent of administrative boundaries while preserving patient confidentiality. Using this approach, the spatial scale can be manipulated to a practically, geographically and computationally sensible scale. It does, however, require individual level geocoded data, which may not be accessible due to confidentiality concerns. Spatial data may also be available at various geographical scales and hence there is a need to combine information from multiple sources (Gotway and Young, 2002). Cramb *et al*. (2011a) mapped two health outcome measures in the Atlas, namely the incidence estimates and the relative survival estimates (discussed in the following Section). Incidence is a measure of the risk of developing a disease within a specified period of time. Relative survival is the standard measure of survival from a disease in population-based disease survival studies (Yu, 2013). Each of these outcomes requires specific input data (Appendix Part E, Boxes 4 and 5). Although other estimates of disease, such as prevalence, are beyond the scope of this article, Bayesian mapping approaches are described in Congdon (2006).

Bayesian spatial statistical models

A response variable is the event studied and expected to vary whenever the independent variable is altered. It is also known as a dependent variable. Here we consider two response variables, namely the number of cancers diagnosed (incidence model) and the number of deaths within x years of diagnosis (relative survival model). Because both response distributions are counts, and the disease is less common, a Poisson distribution is used to model them (Appendix Part E, Box 6).

The resulting estimate for the incidence of a disease is known as the standardised incidence ratio (SIR; Appendix Part D), which is an estimate of relative risk within each area based on the population size, that compares the observed incidence against the expected incidence. The SIR explains if the observed incidence in a particular area is higher or lower than the average across all areas included, given the age and sex distribution and population size of the area.

The relative survival of a disease is modelled using an excess mortality model that contrasts the mortality in the background population with disease mortality. The survival model results in an excess hazard, which is called the relative excess risk (RER). The RER informs the relative survival (Appendix Part D) of a disease within each area, by reporting the risk of death within a certain number of years of diagnosis after adjusting for broad age groups, compared to the average. The SIR and RER are further explained in Appendix Part A.

Small-area disease data typically exhibit spatial correlation due to spatial structure in the unknown risk factors. The presence of spatial correlation can be caused by a combination of socio-demographic clustering and environmental effects (Browning *et al*., 2003). Traditional regression models assume independence of random effects and so ignore the potential presence of spatial correlation. This may lead to false conclusions regarding covariate effects and unstable risk estimates (Fahrmeir and Kneib, 2011).

The spatial correlation can be accounted for using spatial smoothing techniques, by estimating the effect of interest at a location using the effect values at nearby locations (Wang, 2006). Spatial smoothing approaches based on neighbourhood dependence are widely employed in disease mapping where areas with a common boundary are treated as neighbours (Paciorek, 2013). By accounting for the spatial correlation, model inference, prediction and estimation can be improved (Haran, 2011). The effect of the arbitrary geographical boundaries can also be reduced via spatial smoothing. Other smoothing techniques include interpolation methods, kernel regression, kriging and partition methods (Lawson *et al*., 2003; Goovaerts, 2006a).

Two popular ways of defining a neighbourhood structure for the modelling of spatial correlation are the Queen definition and the Rook definition. The Rook method defines that two areas are considered neighbours if they share a common boundary whereas the Queen method specifies that two areas are termed neighbours if they share a common boundary or vertex. Following Earnest *et al*. (2007), the illustration of these two methods for defining a neighbourhood structure is given in Figure 1. Such information can be used to calculate the average of spatially correlated random effects of neighbours for area *i*.

The following Bayesian spatial models take the spatial correlation into account by incorporating spatially correlated random effects. Both the incidence and relative survival models assume a Poisson distribution for the observed data and contain spatial and unstructured (non-spatial) random effects. The well-known Bayesian spatial model of Besag *et al*. (1991) is widely used to model disease incidence (Appendix Part E, Box 7) as it has desirable properties for disease mapping, particularly in modelling the geographical dependence between neighbouring areas (Best *et al*., 2005). The incidence model can also be used to model mortality. With regard to relative survival, the excess mortality can be modelled via a generalised linear model, using exact survival times (Dickman *et al*., 2004). The excess mortality is the mortality that is attributable to a particular disease. It is a measure of the deaths, which occur over and above those that would be expected for a given population. Such a Bayesian relative survival model (Appendix Part E, Box 8) has been used by Fairley *et al*. (2008) and Cramb *et al*. (2011a). See Appendix Part A for the statistical models for incidence and relative survival. In both models, the spatial random effect is the component that accounts for spatial correlation between neighbouring areas. The unstructured or non-spatial random effect accounts for the unexplained variation in the model.

In a Bayesian analysis, it is assumed that all parameters arise from a probability distribution. As such, distributions representing the likely spread of values are placed on each parameter. Commonly, a vague Normal distribution such as one with mean 0 and variance 1.0×10^6 or Normal $(0,1.0\times10^6)$ is used for the intercept or coefficients of predictor terms (Appendix Part D). Vague priors refer to distributions with high spread, such as a Normal distribution with extremely large variance. Such a distribution gives similar prior value over a large range of parameter values.

Generally, the unstructured (non-spatial) random effects and the spatial random effects are both assigned a prior distribution with additional hyperparameters (Appendix Part E, Box 9). To allow for spatial correlation, commonly an intrinsic conditional autoregressive (CAR) distribution is used. The CAR prior models the spatial dependence in a study region by effectively borrowing information from neighbouring areas than from distant areas and smoothing local rates toward local, neighbouring values. The method provides some shrinkage and spatial smoothing of the raw relative risk estimates (Clayton and Kaldor, 1987). This results in a more stable estimate of the pattern of the underlying disease risk than that provided by the raw estimates. Consequently, the variance in the associated estimates is reduced and the spatial effect of geographical differences can be identified. This prior has been widely employed in disease mapping to study the geographical variation of disease risk (Clayton and Bernardinelli, 1992; Mollié, 1996; Wakefield *et al*., 2000), and works particularly well to smooth out variability not relevant to the underlying risk (Assunção and Krainski, 2009).

Commonly, both the precision (inverse of the variance) hyperparameters (Appendix Part D) are assigned a Gamma distribution. Alternative hyperprior distributions may include placing either a Uniform or half-Normal distribution on the standard deviation (square root of the variance) (Gelman and Hill, 2006).

The prior distributions used for the parameters may influence the results and therefore should be carefully considered and compared. There are two issues to consider when deciding on a prior distribution (Gelman, 2002): i) what information is going into the prior distribution; and ii) the impact on the resulting posterior distribution. A sensitivity analysis (Appendix Part D) (Junaidi *et al*., 2011) can be used to investigate the dependence of the posterior distribution on prior distributions by comparing posterior inferences under different reasonable choices of prior distribution. A literature review is usually helpful to determine the prior distributions being used in similar Bayesian models.

Computation

The complexity of these models means they cannot be solved analytically. Instead, some method of approximation is required. One approach is to use Markov chain Monte Carlo (MCMC) methods (Appendix Part D), which samples from the posterior distribution. A variety of software is available to conduct MCMC, including BUGS (Bayesian inference Using Gibbs Sampling), JAGS (Just Another Gibbs Sampler), Stan and BACC (Bayesian Analysis, Computation & Communication). WinBUGS is one of the most popular options (Brooks *et al*., 2011) that provides great flexibility in Bayesian modelling, has a simple programming language (Crainiceanu *et al*., 2005) and interfaces with multiple statistical software, including R, Matlab, Stata and SAS. See Appendix Part B for the WinBUGS code for the discussed models. Some useful resources to help learn WinBUGS include Lawson *et al*. (2003), Lunn *et al*. (2012), Ntzoufras (2009), Lykou and Ntzoufras (2011), and Spiegelhalter *et al*. (2003). Bayesian computation for the above models can also be conducted in R (R Core Team, 2012), by call-

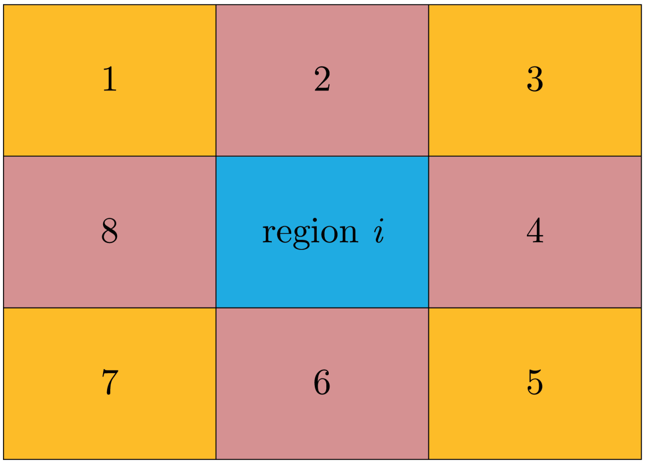

Figure 1. The representation of neighbourhood structure of area *i*. Based on the Rook method, neighbours for area *i* include areas 2, 4, 6 and 8, while the Queen method defines regions 1-8 as neighbours of area *i*.

ing the INLA programme and adopting the integrated nested Laplace approximation (INLA) approach proposed by Rue *et al*. (2009). The INLA approach performs Bayesian inference for spatial models and is able to return accurate parameter estimates in a much shorter time than MCMC. The use of R-INLA for statistical analysis in various disciplines is increasingly common in recent years, including disease mapping. Appendix Part C provides R-INLA code to perform computation for the discussed models. Some useful resources for getting started with R-INLA include Schrödle and Held (2011a, 2011b), Blangiardo *et al*. (2013), and Rue *et al*. (2012). To incorporate neighbourhood dependence into the Bayesian models, a neighbourhood matrix is required. The neighbourhood matrix contains a list of neighbours for an area. Freely available software programmes that will calculate a neighbourhood matrix include GeoDa (Anselin *et al*., 2006), the spdep R package Bivand *et al*. (2011), or within WinBUGS.

Making decisions

Perhaps the greatest advantage of Bayesian methods is the diversity of options available to assist in the decision making process. Communicating results in a way that is easily interpretable and accurate enables informed decisions to be made. Here we outline some of the ways modelled estimates can be used and visualised.

The SIR and RER estimates produced using the methods described in the previous sections are two commonly seen measures of disease risk. Appendices B and C outline the code required for producing the estimates. The estimates produced by Bayesian models give great flexibility in reporting results, including comparison of the risk estimates against the average, ranking estimates, and/or examining the uncertainty around the estimates.

Ranking of disease estimates ensures that public health investigations or interventions are prioritised correctly (Shen and Louis, 2000). In the Bayesian context, the posterior distributions of health outcome measures (such as SIR and RER) allow for the calculation of rank estimates of each area (Clayton and Kaldor, 1987; Lawson *et al*., 2000). For instance, Athens *et al*. (2013) use five health outcome measures to obtain county rank estimates for a composite health outcome measure. The five health outcome measures are converted to a score, and then ranked by weighted means. The ranking of health outcomes is useful for representing health performance of each area which can then be used to inform health decision making.

Moreover, comparison between two areas can be made easily in the Bayesian framework. Outside of Bayesian methods, it may be difficult and problematic to conduct a large number of pairwise comparisons for all areas using post-hoc tests (Jaccard *et al*., 1984). The problem is that by conducting so many comparisons, the probability of finding some of the differences statistically significant by chance alone increases. The Bayesian context eliminates this issue with pairwise comparisons of the posterior distributions.

Bayesian methods produce measures of uncertainty for each modelled estimate. The uncertainty attached to the spatial distribution of risk values across the study region can be known as spatial uncertainty (Goovaerts, 2006b). It is valuable to visualise spatial uncertainty as it provides local details of the spatial variation of the risk, as well as an input to resource allocation, management and policy strategies. Several methods have been proposed to describe the uncertainty attached to the smoothed rates, including mapping the 95% credible interval (Appendix Part D) of the posterior distribution of smoothed rates (Johnson, 2004) and the probability that the risk in each small area exceeds a certain threshold (Richardson *et al*., 2004).

Under the Bayesian paradigm, there is great flexibility in communicating and visualising results. Options include maps or graphs of the smoothed estimates, their associated uncertainty, or the probabilities

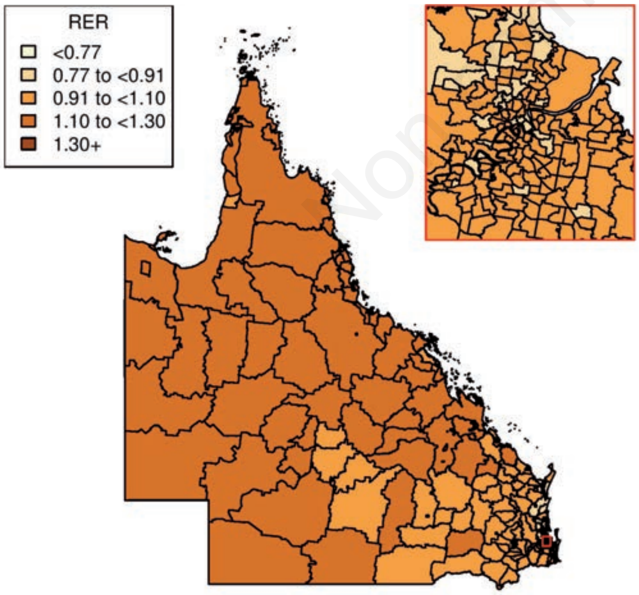

Figure 2. Bayesian smoothed estimate of relative excess risk (RER). To show the spatial pattern of the underlying risk, the median of the posterior distribution of statistical local area (SLA)-level RER is mapped. An inset of South-East Queensland is provided for greater detail as this region has a large number of SLAs. Thematic categories are based on fixed breaks method.

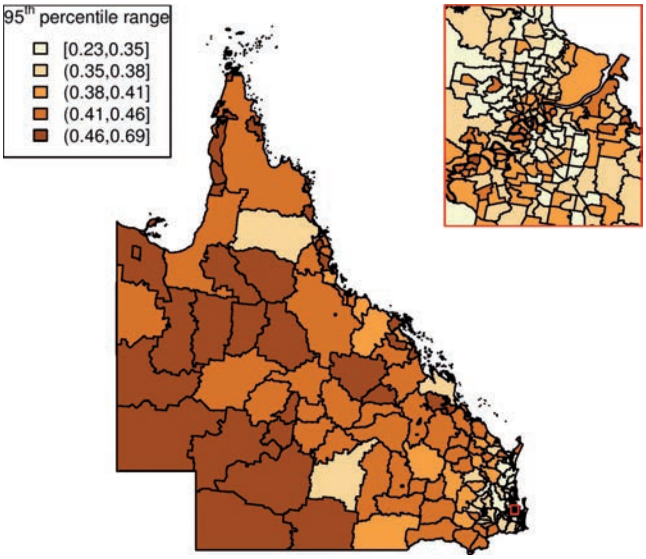

Figure 3. Uncertainty of Bayesian smoothed estimate of relative excess risk (RER). This map depicts the uncertainty associated with the estimates of relative risk. The 95th percentile range (97.5th minus the 2.5th percentile) of the 10,000 values sampled from the posterior distribution of RER for each statistical local area (SLA) is mapped here. An inset of South-East Queensland is provided for greater detail as this region has a large number of SLAs. Thematic categories are based on quintiles.

of being above/below certain values. Mapping of disease rates or outcomes facilitates comparison of spatial patterns in disease rates between males and females, between age groups, between races, over time, and motivates comparison with patterns of potential causes (Brewer and Pickle, 2002). By comparing disease rates of different areas, clues to possible causation may be found and this serves as a starting point for further investigation.

The purpose of this Section is to showcase various visualisations that can be produced using the outputs obtained from Bayesian modelling techniques and the associated interpretation. This is demonstrated on a common cancer with poor survival: male lung cancer in Queensland. Figures 2 to 7 present an array of maps or plots based on the results from modelled survival (RER of death within 5 years of diagnosis) for each SLA that are useful for communicating the results of statistical analysis via the Bayesian paradigm. The RER expresses the risk of cancer patients dying from their cancer within five years of diagnosis in an SLA compared to the Queensland average (RER=1), and therefore should not be directly compared between two SLAs. The figures were produced using the software R, package maptools.

Figure 2 maps the posterior distribution of SLA-level RER and provides a picture of the spatial pattern of the underlying risk. Figure 3 depicts the uncertainty associated with the Bayesian estimates of RER by mapping the 95th percentile range of the 10,000 values sampled from the posterior distribution of RER for each SLA. A graph showing the ranked RER with the associated 95% credible interval for each SLA is provided in Figure 4. Horizontal box plots (Appendix Part D) of the RER estimates by socioeconomic status and rurality are provided in Figure 5 to provide additional information about where the extent of variability across the Queensland state. Figure 6 maps the SLAs having a 90% probability of RER being higher than the Queensland average (RER=1) (highlighted in red) and the SLAs having at least a 90% probability of RER being lower than the Queensland average (RER=1) (highlighted in blue). Figure 7A depicts the probability of the SLAs having RER exceeding 1 and Figure 7B depicts the probability of the SLAs having RER exceeding 1.2.

Results and Discussion

In this article we have outlined the benefits of Bayesian models for both analysis and visualisation. The public health arena regularly makes practical decisions affecting people's health. To facilitate decisions, it is vital that the analysis is conducted appropriately, and results are communicated effectively.

Bayesian methods are increasingly being used to analyse routinely collected data. The Bayesian framework is now the tool of choice in many applied statistical areas, including disease mapping (Lawson *et al.*, 1999). In small area studies, Bayesian methods often have better model fit than non-Bayesian smoothing methods (Lawson *et al.*, 2000). Greater flexibility in distributional assumptions is possible under Bayesian methods than in traditional regression models (Waller and Gotway, 2004). Whether to standardise response rates depends on the study objectives. For the cancer atlas, it was desirable to remove the influence of age, so that differences were not due to different age structures between areas. For incidence, we used the SIR, which adjusts for the area-specific age and sex structure. An alternative method to standardisation for dealing with confounders is via the use of regression models (McNamee, 2005). These can be particularly useful when multiple confounders need to be controlled for simultaneously. For relative survival, we included age in the regression equation to remove its

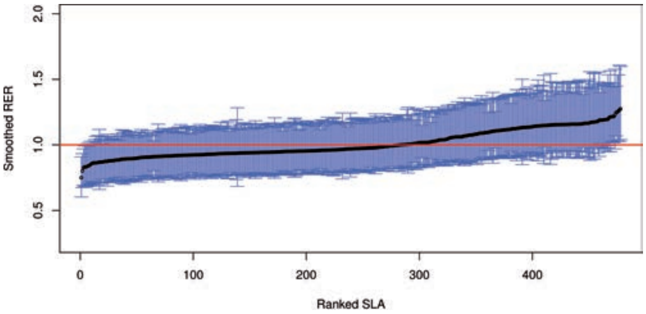

Figure 4. Uncertainty of Bayesian smoothed estimate of relative excess risk (RER). The 95% credible interval (97.5th-2.5th percentile) of the 10,000 values sampled from the posterior distribution of RER for each statistical local area (SLA) is plotted here. This plot shows how much reliance can be placed on the estimates. The black line is the median RER for each SLA. The blue vertical lines are the 95% credible intervals, and indicate the amount of uncertainty associated with each estimate. The red line shows the Queensland average (set to 1).

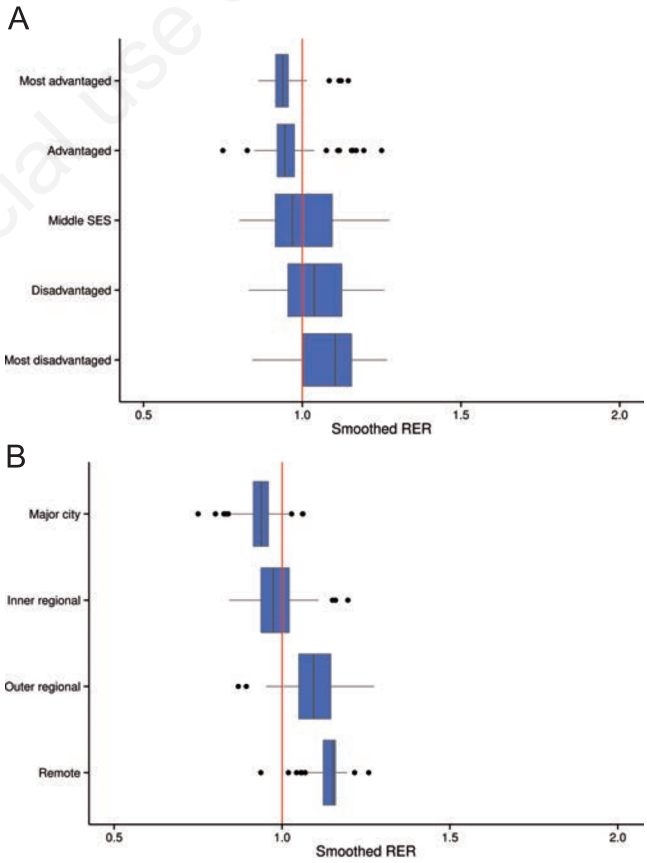

Figure 5. Distribution of smoothed relative excess risk (RER) estimates according to: socioeconomic status (A) and rurality (B). The distributional plots reflect the general patterns in the smoothed RER estimates across the area-based categories of socioeconomic status and rurality. These plots show the proportion of RER estimates that are above or below the Queensland average (vertical red line) within each of the area-based categories. The plots only present the range of point estimates, and so do not take the amount of uncertainty associated with each statistical local area-specific estimate into account.

influence on the results. However, if the purpose of a study is to identify where the highest rates of disease are, such as for service provision, then there is no need to standardise (or otherwise adjust) the incidence rates. This is because the cause of the variation (whether sex, age or other factors), is inconsequential.

Visualising disease patterns through maps remains an effective method to convey a large amount of information in an engaging way. Few modern day visualisations include uncertainty measures, yet this greatly assists in decision-making. Online, interactive visualisations can dynamically link maps (*e.g.* Figure 2 showing the smoothed Bayesian RER), with plots of the uncertainty (*e.g.* Figure 3 showing the 95% credible interval for each area). Selecting an area would then highlight the corresponding region in both plots, providing much greater information to the user.

There are limitations associated with using routinely collected data. Determining the direction of causation may not be possible. Often there is a lag time between exposure and disease detection, and patients may move during this time. Bayesian methods also have certain limitations, including greater computational time if using Markov chain Monte Carlo approaches, and requiring sensitivity analyses to ensure priors are not exerting undue effect. With regard to computation using R-INLA, models must be expressible in the linear model format and there are restrictions on the types of prior distributions that can be assumed. However, we believe the advantages outlined in this article outweigh any limitations. Routinely collected data exist to enable disease monitoring and control. Appropriate analyses convert this data into information, which once communicated, enables action. Bayesian methods not only enable appropriate analyses to be performed, they also provide greater flexibility in visual communications. Can descriptive studies really influence government policy? The disparities identified in the cancer atlas resulted in the Queensland government including a specific objective aimed at reducing the geographic disparities in cancer outcomes in their Strategic Directions (Statewide Health Service Strategy and Planning Unit, 2014). Results were also used in lobbying to increase the amount of financial assistance the government provided to remote patients to offset travel and accommodation costs while obtaining treatment away from home, and the amount provided was subsequently increased. Our experience is that routinely collected data, when appropriately analysed and communicated, facilitate appropriate government action.

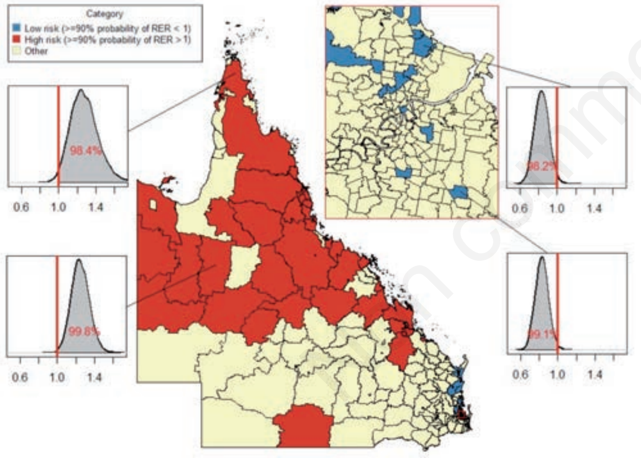
Category
Low risk (>=90% probability of RER < 1)
High risk (>=90% probability of RER > 1)
Other
98.4%
99.8%
98.2%
99.1%
0.6
1.0
1.4

Figure 6. In the Bayesian paradigm, the statistical local areas (SLAs) highlighted in red have a 90% probability of relative excess risk (RER) being higher than the Queensland average (RER=1). This means that the lower 10th percentile of the posterior distribution of RER exceeds 1. The SLAs highlighted in blue express at least a 90% probability of RER being lower than the Queensland average (RER=1). This means that the upper 90th percentile of the posterior distribution of RER is less than 1. The density plots show the posterior distribution of RER for four randomly chosen SLAs where the x-axis is the RER values. The two density plots on the left show that there is more than 90% chance for the RER to be higher than 1. The two density plots on the right show that there is more than 90% chance for the RER to be lower than 1. The percentage of low risk or high risk for each SLA is also given in each density plot. An inset of South-East Queensland is provided for greater detail as this region has a large number of SLAs.

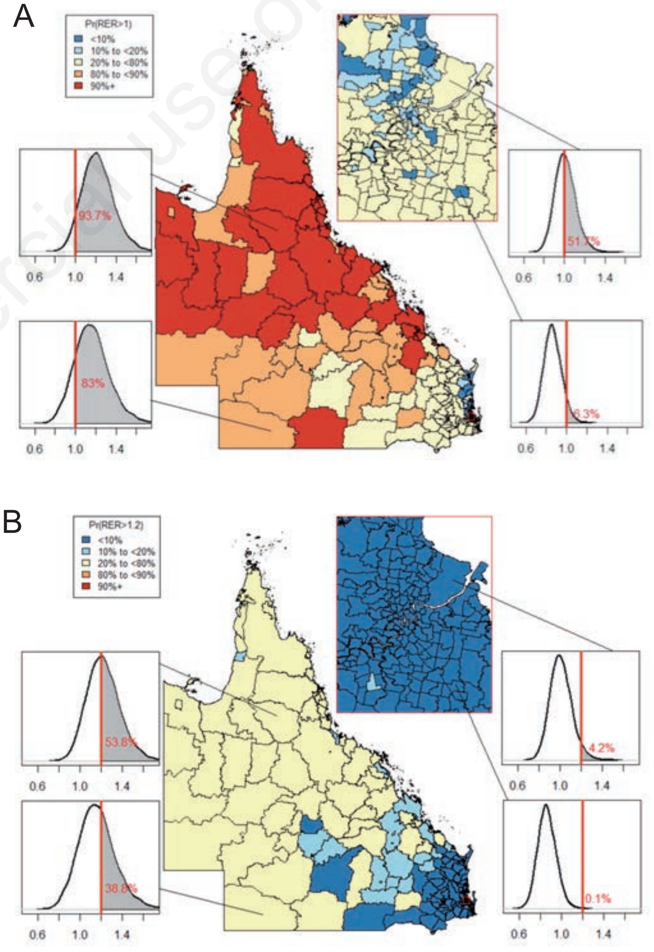

Figure 7. Thematic map depicting the probability of relative excess risk exceeding 1 (A) and 1.2 (B). The threshold 1.2 was chosen to reflect high risk as it lies in the fifth quintile. Four statistical local areas (SLAs) are chosen to demonstrate how the probabilities change when the thresholds change. An inset of South-East Queensland is provided for greater detail as this region has a large number of SLAs.

Conclusions

We hope this article will enable greater understanding and potentially uptake of Bayesian methods in disease mapping, along with available options for communicating estimates and their uncertainty.

References

AIHW, 2014. Cancer in Australia: an overview, 2014. Australian Institute of Health and Welfare, Canberra, Australia.

Ancelet S, Abellan JJ, Vilas VJDR, Birch C, Richardson S, 2012. Bayesian shared spatial-component models to combine and borrow strength across sparse disease surveillance sources. Biom J 54:385-404.

Anselin L, Syabri I, Kho Y, 2006. Geoda: an introduction to spatial data analysis. Geogr Anal 38:5-22.

ApSimon HM, Warren RF, Kayin S, 2002. Addressing uncertainty in environmental modelling: a case study of integrated assessment of strategies to combat long-range transboundary air pollution. Atmos Environ 36:5417-26.

Assunção R, Krainski E, 2009. Neighborhood dependence in Bayesian spatial models. Biom J 51:851-69.

Athens JK, Catlin BB, Remington PL, Gangnon RE, 2013. Using empirical Bayes methods to rank counties on population health measures. Available from: http://www.cdc.gov/pcd/issues/2013/13_0028.htm

Banerjee S, Carlin BP, Gelfand AE, 2014. Hierarchical modeling and analysis for spatial data. 2nd ed. Chapman and Hall/CRC, Boca Raton, FL, USA.

Bernadinelli L, Pascutto C, Best NG, Gilks WR, 1997. Disease mapping with errors in covariates. Stat Med 16:741-52.

Bernardo JM, Smith AFM, 2009. Bayesian theory. Vol 405. John Wiley & Sons Ltd, Hoboken, NJ, USA.

Besag J, York J, Mollié A, 1991. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 43:1-20.

Best N, Richardson S, Thomson A, 2005. A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res 14:35-59.

Bivand R, Anselin L, Berke O, Bernat A, Carvalho M, Chun Y, Dormann CF, Dray S, Halbersma R, Lewin-Koh N, 2011. Spdep: spatial dependence: weighting schemes, statistics and models. Available from: https://cran.r-project.org/web/packages/spdep/index.html

Blangiardo M, Cameletti M, Baio G, Rue H, 2013. Spatial and spatio-temporal models with R-INLA. Spat Spatiotemporal Epidemiol 7:39-55.

Borrell C, Marí-Dell'Olmo M, Serral G, Martínez-Beneito M, Gotsens M, 2010. Inequalities in mortality in small areas of eleven Spanish cities (the multicenter MEDEA project). Health Place 16:703-11.

Brewer CA, Pickle L, 2002. Evaluation of methods for classifying epidemiological data on choropleth maps in series. Ann Assoc Am Geogr 92:662-81.

Brooks S, Gelman A, Jones G, Meng XL, 2011. Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC, Boca Raton, FL, USA.

Browning CR, Cagney KA, Wen M, 2003. Explaining variation in health status across space and time: implications for racial and ethnic disparities in self-rated health. Soc Sci Med 57:1221-35.

Burrough PA, McDonnell R, 1998. Principles of geographical information systems. Oxford University Press, Oxford, UK.

Catelan D, Biggeri A, 2010. Multiple testing in disease mapping and descriptive epidemiology. Geospat Health 4:219-29.

Clayton D, Bernardinelli L, 1992. Bayesian methods for mapping disease risk. In: Elliott P, Cuzick J, English D, Stern R, eds. Geographical and environmental epidemiology: methods for small area studies. Oxford University Press, Oxford, UK, pp 205-20.

Clayton D, Kaldor J, 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43:671-81.

Congdon P, 2006. Estimating diabetes prevalence by small area in England. J Public Health 28:71-81.

Crainiceanu CM, Ruppert D, Wand MP, 2005. Bayesian analysis for penalized spline regression using WinBUGS. J Stat Softw 14:1-24.

Cramb SM, Mengersen KL, Baade PD, 2011a. Atlas of cancer in Queensland: geographical variation in incidence and survival, 1998-2007. Viertel Centre for Research in Cancer Control, Cancer Council Queensland, Brisbane, Australia.

Cramb SM, Mengersen KL, Baade PD, 2011b. Developing the atlas of cancer in Queensland: methodological issues. Int J Health Geogr 10:9.

Cressie N, Chan NH, 1989. Spatial modeling of regional variables. J Am Stat Assoc 84:393-401.

Dickman PW, Sloggett A, Hills M, Hakulinen T, 2004. Regression models for relative survival. Stat Med 23:51-64.

Earnest A, Morgan G, Mengersen K, Ryan L, Summerhayes R, Beard J, 2007. Evaluating the effect of neighbourhood weight matrices on smoothing properties of conditional autoregressive (CAR) models. Int J Health Geogr 6:54.

Elliot P, Wakefield JC, Best NG, Briggs DJ, 2000. Spatial epidemiology: methods and applications. Oxford University Press, Oxford, UK.

English PB, 2001. An introductory guide to disease mapping. Am J Epidemiol 154:881-2.

Ernst J, Zenger M, Schmidt R, Schwarz R, Brähler E, 2010. Medical and psychosocial care needs of cancer patients: a systematic review comparing urban and rural provisions. Deut Med Wochenschr 135:1531-7.

Fahrmeir L, Kneib T, 2011. Bayesian smoothing and regression for longitudinal, spatial and event history data. Oxford University Press, Oxford, UK.

Fairley L, Forman D, West R, Manda S, 2008. Spatial variation in prostate cancer survival in the Northern and Yorkshire region of England using Bayesian relative survival smoothing. Brit J Cancer 99:1786-93.

Gelman A, 2002. Prior distribution. In: El-Shaarawi AH, Piegorsch WW, eds. Encyclopedia of environmetrics. John Wiley & Sons Ltd, Chichester, UK, pp 1634-7.

Gelman A, Hill J, 2006. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York, NY, USA.

Ghosh M, Natarajan K, Waller LA, Kim D, 1999. Hierarchical Bayes GLMs for the analysis of spatial data: an application to disease mapping. J Stat Plan Inference 75:305-18.

Goovaerts P, 2006a. Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. Int J Health Geogr 5:52.

Goovaerts P, 2006b. Geostatistical analysis of disease data: visualization and propagation of spatial uncertainty in cancer mortality risk using Poisson kriging and p-field simulation. Int J Health Geogr 5:7.

Gotway CA, Young LJ, 2002. Combining incompatible spatial data. J Am Stat Assoc 97:632-48.

Green PJ, and S Richardson, 2002. Hidden markov models and disease mapping. J Am Stat Assoc 97:1055-70.

Gurrin LC, Kurinczuk JJ, Burton PR, 2000. Bayesian statistics in medical research: an intuitive alternative to conventional data analysis. J Eval Clin Pract 6:193-204.

Haran M, 2011. Gaussian random field models for spatial data. In: Brooks SP, Gelman A, Jones GL, Meng XL, eds. Handbook of Markov Chain Monte Carlo. CRC Press, Boca Raton, FL, USA, pp 449-78.

IARC, 2014. World cancer report 2014. International Agency for Research on Cancer-World Health Organization, Geneva, Switzerland.

Jaccard J, Becker MA, Wood G, 1984. Pairwise multiple comparison procedures: a review. Psychol Bull 96:589.

Johnson GD, 2004. Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling. Int J Health Geogr 3:29.

Junaidi, Stojanovski E, Nur D, 2011. Prior sensitivity analysis for a hierarchical model. In: Proceedings of the Fourth Annual ASEARC Conference, 17-18 February 2011, University of Western Sydney, Paramatta, Australia.

Kang SY, McGree J, Baade P, Mengersen K, 2014. An investigation of the impact of various geographical scales for the specification of spatial dependence. J Appl Stat 41:2515-38.

Koch T, 2011. Disease maps: epidemics on the ground. University of Chicago Press, Chicago, USA.

Kulldorff M, Song C, Gregorio D, Samociuk H, DeChello L, 2006. Cancer map patterns: are they random or not? Am J Prev Med 30:37-49.

Lawson AB, 2001. Statistical methods in spatial epidemiology. Wiley, Chichester, UK.

Lawson AB, 2009. Bayesian disease mapping: hierarchical modeling in spatial epidemiology. CRC Press, Boca Raton, FL, USA.

Lawson AB, Biggeri AB, Böhning D, Lesaffre E, Viel J-F, Bertollini R, 1999. Disease mapping and risk assessment for public health. John Wiley & Sons, Chichester, UK.

Lawson AB, Biggeri AB, Böhning D, Lesaffre E, Viel J-F, Clark A, Schlattmann P, Divino F, 2000. Disease mapping models: an empirical evaluation. Stat Med 19:2217-41.

Lawson AB, Browne WJ, Rodeiro CV, 2003. Disease mapping with WinBUGS and MLwiN. Vol 11. John Wiley & Sons, Chichester, UK.

Lawson AB, Williams FLR, 2001. An introductory guide to disease mapping. John Wiley & Sons, Chichester, UK.

Lee D, 2011. A comparison of conditional autoregressive models used in Bayesian disease mapping. Spat Spatiotemporal Epidemiol 2:79-89.

Li Y, Brown P, Gesink DC, Rue H, 2012a. Log Gaussian Cox processes and spatially aggregated disease incidence data. Stat Methods Med Res 21:479-507.

Li Y, Brown P, Rue H, al Maini M, Fortin P, 2012b. Spatial modelling of lupus incidence over 40 years with changes in census areas. J Roy Stat Soc C Appl Stat 61:99-115.

López-Abente G, Aragonés N, García-Pérez J, Fernández-Navarro P, 2014. Disease mapping and spatio-temporal analysis: importance of expected-case computation criteria. Geospat Health 9:27-35.

Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D, 2012. The BUGS book: a practical introduction to Bayesian analysis. Chapman and Hall/CRC Press, Boca Raton, FL, USA.

Lykou A, Ntzoufras I, 2011. WinBUGS: a tutorial. Wiley Interdiscip Rev Comput Stat 3:385-96.

Marshall RJ, 1991. A review of methods for the statistical analysis of spatial patterns of disease. J Roy Stat Soc A Sta 154:421-41.

Mason TJ, McKay FW, Hoover R, Blot WJ, Fraumeni JF, 1975. Atlas of cancer mortality for U.S. counties: 1950-1969. US Govt. Printing Office, Washington, DC, USA.

McNamee R, 2005. Regression modelling and other methods to control confounding. Occup Environ Med 62:500-6.

Mollié A, 1996. Bayesian mapping of disease. In: Gilks WR, Richardson S, Spiegelhalter DJ, eds. Markov Chain Monte Carlo in practice. Chapman & Hall, London, UK, pp 359-79.

Ntzoufras I, 2009. Bayesian modeling using WinBUGS. John Wiley & Sons, Hoboken, NJ, USA.

Paciorek CJ, 2013. Spatial models for point and areal data using Markov random fields on a fine grid. Electron J Stat 7:946-72.

R Core Team, 2012. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Richardson S, Thomson A, Best N, Elliott P, 2004. Interpreting posterior relative risk estimates in disease-mapping studies. Environ Health Persp 112:1016.

Rue H, Martino S, Chopin N, 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J Roy Stat Soc B Met 71:319-92.

Rue H, Martino S, Lindgren F, 2012. The R-INLA project. Available from: http://www.r-inla.org

Rytkönen MJ, 2004. Not all maps are equal: GIS and spatial analysis in epidemiology. Int J Circumpolar Health 63:9-24.

Schrödle B, Held L, 2011a. A primer on disease mapping and ecological regression using INLA. Computation Stat 26:241-58.

Schrödle B, Held L, 2011b. Spatiotemporal disease mapping using INLA. Environmetrics 22:725-34.

Shen W, Louis TA, 2000. Triple-goal estimates for disease mapping. Stat Med 19:2295-308.

Spiegelhalter D, Thomas A, Best N, Lunn D, 2003. WinBUGS user manual. Available from: www.mrc-bsu.cam.ac.uk/wp-content/uploads/manual14.pdf

Statewide Health Service Strategy and Planning Unit, 2014. Cancer care services statewide health service strategy 2014. Statewide Health Service Strategy and Planning Unit, Brisbane, Australia.

Thomas DC, 2014. Statistical methods in environmental epidemiology. Oxford University Press, Oxford, UK.

Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A, 2015. Global cancer statistics, 2012. CA: Cancer J Clin 65:87-108.

Wakefield J, 2007. Disease mapping and spatial regression with count data. Biostatistics 8:158-83.

Wakefield JC, Best NG, Waller LA, 2000. Bayesian approaches to disease mapping. In: Elliot P, Wakefield JC, Best NG, Briggs DJ, eds. Spatial epidemiology: methods and applications. Oxford University Press, Oxford, UK, pp 104-27.

Waller LA, Gotway CA, 2004. Applied spatial statistics for public health data. John Wiley & Sons, Hoboken, NJ, USA.

Wang F, 2006. Quantitative methods and applications in GIS. CRC Press, Boca Raton, FL, USA.

Wilkinson D, Cameron K, 2004. Cancer and cancer risk in South Australia: what evidence for a rural-urban health differential? Aust J Rural Health 12:61-6.

Woods LM, Rachet B, Coleman MP, 2006. Origins of socio-economic inequalities in cancer survival: a review. Ann Oncol 17:5-19.

Yu B, 2013. A class of transformation covariate regression models for estimating the excess hazard in relative survival analysis. Am J Epidemiol 177:708-17.