20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

The impact of spatial scales on discretised spatial point patterns

Su Yun Kang [a b], **James McGree** [a b], **Kerrie Mengersen** [a b]

[a] *Mathematical Sciences School, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia*
[b] *CRC for Spatial Information, 204 Lygon Street, Carlton, Victoria 3053, Australia*
Email: s7.kang@qut.edu.au

Abstract: Spatial data are common in health sciences and are available at various spatial scales such as the point, grid or area level. This research considers modelling of point level data, which in practice could resemble disease data with exact residential locations, by discretizing the study region into regular grid cells. Modelling of health data at the grid level is desirable as it is geographically more accurate than using area level data and yet protects patient confidentiality. The challenge is to specify an appropriate spatial scale for discretization of point patterns. We investigate how changes in grid cell size affect model outcomes for various structures of spatial point patterns. A Bayesian spatial model is used to evaluate the impact of varying spatial scales on model outcomes. Estimation is based on a Bayesian spatial smoothness prior to model spatial dependence of neighboring grid cells, namely an intrinsic Gaussian Markov random field (IGMRF). Bayesian computation is carried out using integrated nested Laplace approximation (INLA). The impact of varying spatial scales is studied in a simulation study. The simulated data consist of various spatial patterns that resemble different patterns of point level health data in realistic settings, including inhomogeneous point patterns, patterns with local repulsion, patterns with local clustering, and patterns with local clustering in the presence of a larger-scale inhomogeneity. The evaluation criteria used in this study include the spatial correlation coefficient, the coefficient of variation of the spatially structured effect, and the mean squared error between the observed counts and the estimated counts. Based on the results, we note that complicated spatial patterns such as inhomogeneous point patterns and spatially clustered patterns tend to be more sensitive to the changing spatial scales, compared to homogeneous point patterns. It is therefore recommended to repeat the spatial analyses at multiple spatial scales in order to determine the best scale to analyze the data in order to address the inferential aims of interest. In particular, it is noted that fine grid cell sizes do not necessarily improve inferential outcomes as there has to be sufficient information in the grid cells.

Keywords: *Grid level modelling, integrated nested Laplace approximation, intrinsic Gaussian Markov random field, spatial scale*

1 INTRODUCTION

Spatial data are common in health sciences and are available at various spatial scales such as the point, grid or area level. In the context of health sciences, area level data are widely available and commonly utilized for convenience. For instance, an event of interest is often aggregated to administrative districts in order to protect patient confidentiality. Despite the popularity of modelling health data at the area level, aggregation of data increases spatial correlation (Song et al., 2011). Other concerns raised about aggregated data include bias in estimates due to ecological fallacy (Robinson, 1950), loss of information, and issues of overlapping boundaries and artificiality of administrative or political boundaries (Louie and Kolaczyk, 2006; Kirby, 1996). Ecological fallacy refers to the difference between individual and group level estimates of risk measures. In small area modelling, summary statistics collected for the group are used to make inference about the nature of individuals within that group. However, summary statistics that describe group features do not necessarily hold for individuals within that group. When a relationship observed at the group level is assumed to apply at the individual level, the fallacy is committed.

Despite these acknowledged problems, the history of disease mapping has shown that point level data has received less attention than area level data modelling. This is due in part to difficult access to individual level data for confidentiality and privacy reasons, lack of geocoding of disease outcomes, concerns about the impact of spatial misalignment, missing data and other issues. On the one hand, models based on individual level data are, however, able to uncover local-level inequalities frequently masked by health estimates from large areas such as states, regions or cities (Borrell et al., 2010). On the other hand, modelling of point level data is computationally demanding when large datasets are involved due to dense covariance matrices.

A compromise between these two methods is modelling of spatial data at a grid level, by discretizing the study region into regular grid cells, or by utilizing spatial data which are collected directly at the grid level (also known as raster data). Raster data are especially useful in representing geographic phenomena that vary continuously across space such as population density and other demographic characteristics that are important in health modelling (Chang, 2010; Lai et al., 2009). Despite being less common than area level data modelling, grid level modelling approaches have become increasingly popular in recent years (Baddeley et al., 2010; Li et al., 2012). Grid level modelling of disease data is geographically more accurate than using aggregated data and yet protects patient confidentiality. It allows the spatial scale at which the data are to be modelled to be manipulated to a computationally and practically sensible scale, and avoids the problem of changing geographical boundaries over time that can occur with area level data.

Determining an appropriate spatial scale is known to be a challenge in grid level modelling of spatial data. The same basic data may yield different results when aggregated in different ways. The sensitivity of spatial analyses to the definition of spatial scales is so-called the modifiable areal unit problem (MAUP) (Openshaw and Taylor, 1981) which is widely known in the statistical and geographical literature. Given the acknowledged MAUP, it is of interest to investigate the effect of changing scale on the analysis of spatial pattern and process. At present, little is known about the impact of changing scales on the outcome of spatial models at different spatial patterns. In this paper, we design a simulation study to examine how changes in grid cell size effect model outcomes, in particular model goodness-of-fit at various spatial scales. Bayesian inference is carried out using integrated nested Laplace approximation (INLA) throughout the study. The simulated datasets consist of point data with various spatial patterns such as inhomogeneous point patterns, patterns with local repulsion, patterns with local clustering, and patterns with local clustering in the presence of a larger-scale inhomogeneity.

2 METHODS

Let X be a spatial point-based dataset embedded in an observation window S which is discretized into $n_1 \times n_2$ grid cells $\{s_{ij}\}$ with area $|s_{ij}|$ for $i = 1, ..., n_1$ and $j = 1, ..., n_2$. Let N_{ij} denote the observed number of points in each grid cell s_{ij}. Assume that N_{ij} are conditionally independent Poisson counts

$$N_{ij} \sim \text{Po}(|s_{ij}|\lambda_{ij}),$$

where λ_{ij} denotes the intensity in each grid cell. We are interested in modelling the log-intensity ($\eta_{ij} = \log(\lambda_{ij})$) of the Poisson process. Spatial variation in the log-intensity is modelled using different components including μ, u_{ij}, and v_{ij}, where μ refers to the common intercept term, u_{ij} is a spatially structured term that describes the effect of the location by assuming that geographically close areas are more similar than distant areas, and v_{ij} is an unstructured term that accounts for unexplained variability in the process.

In this model, the spatially structured component, u_{ij}, is assigned an intrinsic Gaussian Markov random field

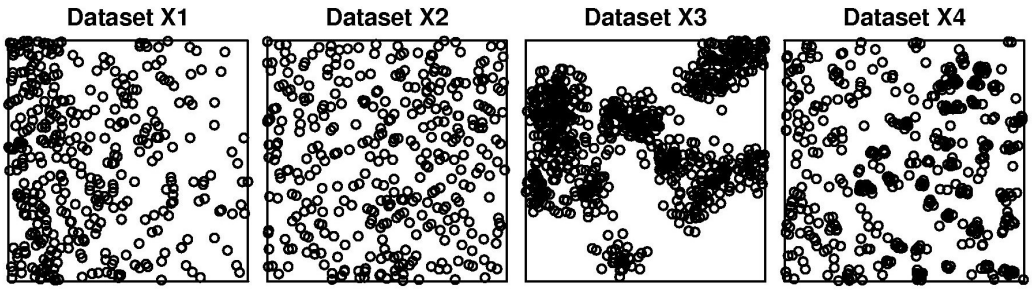

Figure 1. Four patterns of simulated point-based data

(IGMRF) prior with unknown precision (inverse variance) τ_u. The spatially unstructured component, v_{ij}, is assumed to be independent and identically distributed (i.i.d.) and normally distributed with zero mean and unknown precision τ_v; and gamma priors are assigned to the precision parameters τ_u and τ_v.

An IGMRF for u_{ij} is defined as

$$u_{ij}|\boldsymbol{u}_{-ij}, \tau_u \sim \mathcal{N}\left(\frac{1}{n_{ij}}\sum_{ij\sim kl} u_{kl}, \frac{1}{n_{ij}\tau_u}\right),$$

where n_{ij} is the number of neighbors of grid cell s_{ij}, $\boldsymbol{u}_{-ij}$ denotes all elements in $\boldsymbol{u}$ except for u_{ij}, and $ij \sim kl$ indicates that the two grid cells are neighbors that share a common boundary. A sum-to-zero constraint is imposed on u_{ij} to ensure identifiability of the intercept μ. We refer the reader to Besag et al. (1991) and Rue and Held (2005) for further details. This model has been widely applied in disease mapping to study spatial variation of disease risk. However, the neighborhoods in these papers were defined in terms of administrative districts, while here we consider a finer neighborhood structure in terms of (regular) grid cells. The IGMRF model can be written as follows,

$$\text{IGMRF model}: \eta_{ij} = \log(\lambda_{ij}) = \mu + u_{ij} + v_{ij}.$$

In light of the computational cost of Markov chain Monte Carlo (MCMC) methods for spatial inference, we adopt the integrated nested Laplace approximation (INLA) approach proposed by Rue et al. (2009). INLA performs approximate Bayesian inference for latent Gaussian models. Computation in this study is performed in the `R` package, by calling the `inla` program.

3 SIMULATION STUDY

3.1 Description of data

The purpose of this simulation study is to investigate the impact of spatial scales at various spatial structures of point-based data. As guided by Illian et al. (2012), we considered four different situations: inhomogeneous point patterns, patterns with local repulsion, patterns with local clustering, and patterns with local clustering in the presence of a larger-scale inhomogeneity. The inhomogeneous point patterns (dataset **X1**) were generated from an inhomogeneous Poisson process with trend function $\lambda = 1000\exp(-2x)$ on the unit square. For the patterns with local repulsion (dataset **X2**), we generated point-based data from a homogeneous Strauss process, with medium repulsion $\beta = 700$ (intensity parameter), interaction parameter $\gamma = 0.8$ and interaction radius $r = 0.05$, on the unit square. To generate the clustered patterns (dataset **X3**), we simulated a homogeneous Thomas process with parameters $\kappa = 10$ (the intensity of the Poisson process of cluster centers), $\sigma = 0.05$ (the standard deviation of the distance of a point from the cluster center) and $\mu = 50$ (the expected number of points per cluster), on the unit square. For the last spatial pattern (dataset **X4**), we generated the data from an inhomogeneous Thomas process with parameters $\sigma = 0.01$ and $\mu = 5$ and a simple trend function for the intensity of parent points given by $\kappa(x_1, x_2) = 100x_1$, on the unit square. Each pattern was then superimposed with a pattern generated from an inhomogeneous Poisson process with trend function $\lambda = 500\exp(-2x)$. See Figure 1 for illustrations of the point-based data.

3.2 Model fitting and evaluations

We fit the IGMRF model to each dataset which is discretized into $n_1 \times n_2$ regular grid cells. The precision parameters of the spatial effect and unstructured effect are both assigned gamma priors with parameters $(1, 0.01)$ to impose the same level of spatial smoothing on the spatial field throughout the entire simulation study. To illustrate modelling at different spatial scales, we set $n_1 = n_2 = 5, 10, 15, 20, 25, 30, 35, 40, 45, 50$ in the simulation study, resulting in $25, 100, 225, \ldots, 2500$ grid cells, respectively. A neighborhood structure has to be specified to model spatial dependence of grid cells via the IGMRF prior. We use the `cell2nb` function in the `spdep R` package (Bivand et al., 2011) to generate a list of neighbors for the grid cells, by applying a rook definition of neighborhood, where two grid cells are termed neighbors if they share a common edge.

To measure the importance of spatial correlation in the data, the spatial correlation coefficient is calculated at each spatial scale (Flask and Schneider IV, 2013),

$$\phi = \frac{\sigma_u}{\sigma_u + \sigma_v},$$

where σ_u is the standard deviation of the spatially structured effects $\boldsymbol{u}$, and σ_v represents the standard deviation of the unstructured random effects $\boldsymbol{v}$. The spatially structured component becomes increasingly dominant as this statistic approaches unity.

The coefficient of variation of the spatially structured effect is also calculated,

$$\text{CV}_u = \frac{\sigma_u}{\bar{\boldsymbol{u}}},$$

where σ_u is described above and $\bar{\boldsymbol{u}}$ is the mean of the spatially structured effects $\boldsymbol{u}$. The CV_u, a normalized measure of dispersion, shows the variation of the spatial effect in relation to its mean.

In terms of the comparison of predictive performance at various spatial scales, we calculate mean squared error (MSE) between the observed counts in a grid cell, N_{ij}, and the estimated counts in the respective grid cell, $\hat{N}_{ij}$,

$$\text{MSE} = \frac{\sum_{i,j=1}^{n_1 \times n_2} \left(\hat{N}_{ij} - N_{ij}\right)^2}{n_1 \times n_2}.$$

A smaller MSE indicates a better predictive performance.

4 RESULTS

Figure 2 presents the spatial correlation coefficient, ϕ, at various spatial scales for the four datasets. Dataset **X1** which consists of inhomogeneous spatial pattern appears to have ϕ close to 1 at all scales, which suggests that the spatially structured component is dominant over the unstructured effect. The change in grid cell size does not affect ϕ for this spatial pattern. Dataset **X2** (point pattern with local repulsion) produces ϕ that decreases gradually from 0.56 to 0.35 across the changing spatial scales, signifying the diminishing spatial effect as the grid cell size becomes smaller. The spatially clustered pattern in dataset **X3** results in a drastic increase in ϕ from 0.01 at the scale 5×5 to 0.99 at all subsequent scales, suggesting an abrupt dominance of the spatial effect. Dataset **X4** which contains small clusters is also sensitive to the varying spatial scales as ϕ is observed to vary across the scales. The spatial component in this spatial pattern becomes dominant at small grid cell sizes (the scale 35×35 and beyond). The results for all four datasets hence suggest that complicated spatial patterns such as inhomogeneous and clustered patterns appear to be sensitive to the change in spatial scales and produce varying ϕ at different scales.

The coefficient of variation of the spatially structured effect (CV_u) at the different scales is shown in Figure 3. Dataset **X1** results in CV_u that decreases slightly across the changing spatial scales. Dataset **X2** produces CV_u that is rather consistent across the spatial scales. On the other hand, dataset **X3** shows a drastic drop in CV_u from 1.46 at the scale 5×5 to 0.22 at the scale 10×10 which then decreases gradually at all subsequent scales. Dataset **X4** displays a slight increase in CV_u at coarse spatial scales (scales 5×5 to 30×30) which then drops abruptly from 1.43 at the scale 30×30 to 0.14 at the scale 35×35. Based on the results for all datasets, it is observed that modelling at a smaller scale improves the model performance for inhomogeneous and clustered spatial patterns, by reducing the CV_u which is a measure of dispersion normalized by its mean. However, a

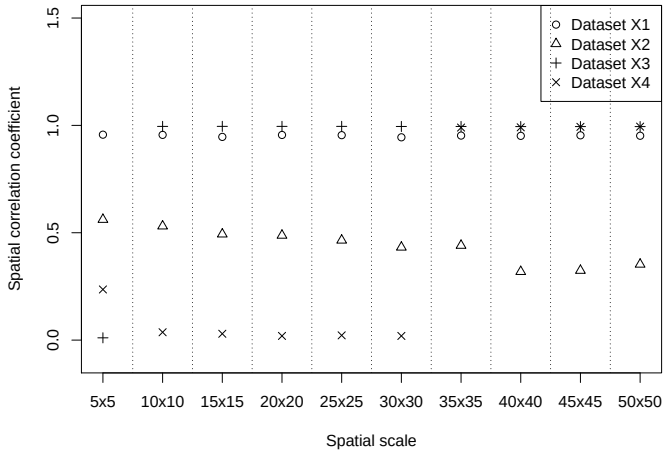

Figure 2. Spatial correlation coefficient for each dataset

Table 1. The MSE at various spatial scales for each dataset

Spatial scale	**Dataset X1**	**Dataset X2**	**Dataset X3**	**Dataset X4**
5×5	2.964	8.021	0.396	8.892
10×10	1.793	2.630	0.390	1.503
15×15	1.184	1.286	0.326	0.848
20×20	0.668	0.762	0.304	0.476
25×25	0.476	0.493	0.279	0.380
30×30	0.367	0.360	0.252	0.266
35×35	0.266	0.265	0.218	0.234
40×40	0.202	0.223	0.210	0.186
45×45	0.159	0.176	0.183	0.149
50×50	0.132	0.140	0.159	0.125

very fine scale is not necessarily favored as it is shown that CV_u becomes consistent after a certain spatial scale for all datasets. Further discretization into finer grid cells does not contribute to significant decrease in CV_u.

In regard to the predictive performance, Table 1 presents the MSE at various spatial scales for all four datasets. It is evident that the MSE reduces across the decreasing grid cell sizes for all datasets, suggesting that the predictive performance of the model is improved at finer scales. For all datasets, it appears that the MSE reduces substantially at coarser scales but decreases gradually at finer scales. For instance, the MSE for dataset **X1** has only slight decrease after the scale 20×20, which means that further discretization into scales finer than this is not necessary. Similarly, the MSE for dataset **X2** becomes rather consistent after the scale 20×20, whereas the MSE for dataset **X3** seems to favor small grid cell sizes. The MSE for dataset **X4** shows that modelling at scales finer than 15×15 does not improve the prediction significantly.

5 CONCLUSIONS

In this study, we investigate the impact of varying spatial scales on the model outcomes using a simulation study. The simulated data consist of various spatial patterns that resemble different patterns of point level health data in realistic, including inhomogeneous point patterns, patterns with local repulsion, patterns with local clustering, and patterns with local clustering in the presence of a larger-scale inhomogeneity. The evaluation criteria used in this study include the spatial correlation coefficient (ϕ), the coefficient of variation of the

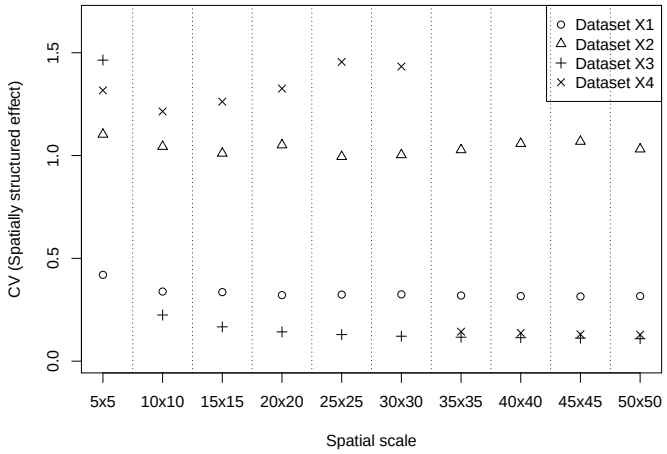

Figure 3. Coefficient of variation of the spatially structured effect for each dataset

spatially structured effect (CV_u), and the mean squared error between the observed counts and the estimated counts (MSE). The results of analyzing all four datasets in this study suggest that different spatial scales may be more applicable for different spatial patterns.

Based on ϕ obtained for each dataset, we note that complicated spatial patterns such as inhomogeneous point patterns and spatially clustered patterns tend to be more sensitive towards the changing spatial scales, compared to homogeneous point patterns. For these spatial patterns, it appears that the dominance of spatial component is largely affected by the size of the grid cells. The CV_u produced in each dataset also supports that inhomogeneous and clustered point patterns are sensitive to the spatial scales. When predictive performance of the model at various spatial scales is of concern, the MSE is used as a criterion for evaluation. As observed in this study, the reducing MSE across the decreasing spatial scales suggests that the predictive performance is improved at finer scales. Given these results, it is recommended to repeat the spatial analyses at multiple spatial scales in order to determine the best scale to analyze the data. One should note that fine grid cell sizes do not necessarily improve model outcomes as there has to be sufficient information in the grid cells.

In this study, we have not included covariates in the model although this is straightforward to do. The inclusion of covariates may alternate the spatial scale effects observed here. The aggregation of counts using grid level modelling approach may still lead to some degree of aggregation or ecological bias when covariates are involved. However, if the grid size is sufficiently small in terms of population and physical characteristics of the exposure, the homogeneity assumption of exposure within areas can be assumed to hold. Thus, there will be less concern about ecological bias when covariates are included in the model.

In the health context, analysis of health data at grid level does not only protects patient confidentiality but also reduces ecological bias as the data are geographically more precise than data aggregated by administrative districts. This approach may be applied to street level data and thus avoid confidentiality issues and the need to obtain patient consent. Finally it is also of interest to consider the inclusion of time components to build spatio-temporal models for discretized spatial patterns. This is a topic for future research.

ACKNOWLEDGEMENT

The work has been supported by the Cooperative Research Centre for Spatial Information, whose activities are funded by the Australian Commonwealth's Cooperative Research Centres Programme.

REFERENCES

Baddeley, A., M. Berman, N. I. Fisher, A. Hardegen, R. K. Milne, D. Schuhmacher, R. Shah, and R. Turner (2010). Spatial logistic regression and change-of-support in Poisson point processes. *Electronic Journal of Statistics 4*, 1151–1201.

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

Bivand, R., L. Anselin, O. Berke, A. Bernat, M. Carvalho, Y. Chun, C. F. Dormann, S. Dray, R. Halbersma, N. Lewin-Koh, et al. (2011). spdep: Spatial dependence: weighting schemes, statistics and models. *R package version 0.5-31, URL http://CRAN. R-project. org/package= spdep*.

Borrell, C., M. Marí-Dell'Olmo, G. Serral, M. Martínez-Beneito, and M. Gotsens (2010). Inequalities in mortality in small areas of eleven Spanish cities (the multicenter MEDEA project). *Health & place 16*(4), 703–711.

Chang, K. (2010). *Introduction to Geographic Information Systems*. McGraw-Hill, New York.

Flask, T. and W. Schneider IV (2013). A bayesian analysis of multi-level spatial correlation in single vehicle motorcycle crashes in ohio. *Safety Science 53*, 1–10.

Illian, J. B., S. H. Sørbye, and H. Rue (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics 6*(4), 1499–1530.

Kirby, R. S. (1996). Toward congruence between theory and practice in small area analysis and local public health data. *Statistics in Medicine 15*(17), 1859–1866.

Lai, P. C., F. M. So, and K. W. Chan (2009). *Spatial Epidemiological Approaches in Disease Mapping and Analysis*. CRC Press, Hoboken.

Li, Y., P. Brown, H. Rue, M. al Maini, and P. Fortin (2012). Spatial modelling of lupus incidence over 40 years with changes in census areas. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(1), 99–115.

Louie, M. M. and E. D. Kolaczyk (2006). A multiscale method for disease mapping in spatial epidemiology. *Statistics in Medicine 25*(8), 1287–1306.

Openshaw, S. and P. J. Taylor (1981). The modifiable areal unit problem. In N. Wrigley and R. Bennet (Eds.), *Quantitative Geography: A British View*, pp. 60–69. Routledge and Kegan Paul, London.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review 15*(3), 351–357.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*, Volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(2), 319–392.

Song, H. R., A. Lawson, et al. (2011). Modeling type 1 and type 2 diabetes mellitus incidence in youth: An application of Bayesian hierarchical regression for sparse small area data. *Spatial and Spatio-temporal Epidemiology 2*(1), 23–33.