STAT0034-Investigating the spatio-temporal variability of spatial point pattern data

Candidate number: PNZV5

October 4, 2021

Contents

1 Introduction **4**

2 Overview of Spatial Point Pattern Data **6**

2.1 Basic description of data 6

2.2 Advantages and Disadvantages of AEGISS Data in Epidemiology Data Analysis 7

3 Exploratory Data Analysis of Spatial Point Pattern **9**

3.1 Intensity in the context of spatial statistics 9

3.2 Adjustment to baseline risk level 10

3.3 General temporal variation in the intensity 11

3.4 Spatial variation in the intensity of spatial point pattern 11

3.5 Spatio-temporal variation in intensity homogeneity 17

4 Spatial Statistical Modelling **19**

4.1 Introduction to common spatial statistical models 19

4.2 Modelling the intensity of AEGISS data 21

4.3 Results 22

5 Discussion **26**

6 Acknowledgement **27**

Abstract

Spatial statistics is a powerful tool in analysing data with spatial structure and is invaluable with many possible real-world applications. This report discusses the result from the analysis of the AEGISS (Ascertainment and Enhancement of Gastrointestinal Infection Surveillance and Statistics) project data on cases of non-specific gastrointestinal infections reported from Hampshire, UK, between 2001 and 2003. Different exploratory analyses demonstrated the inhomogeneity of the AEGISS data and partially show that the inhomogeneity is the consequence of unobserved random effects. The unobserved random effect could not be captured within this analysis due to limits in sources of data. Powers of spatial statistics model as a tool in predicting unobserved cases of infection based on current observation has been partially demonstrated. This report has highlighted the respective edges and drawbacks of spatial statistics as an application-based tool in data analysis.

1 Introduction

Spatial point pattern is a type of data about the spatial information of observations and/or events. Spatial point pattern and spatial data differ from conventional statistical data that the former particularly emphasises the spatial information of data points. Compared with other types of spatial data, spatial point pattern also emphasises that usually each datus consists of a single, volume-less point in space, rather than a fraction of an area (Fig.1). An example of spatial point pattern could be addresses of 1000 new cases of Covid-19 in a London city borough within the last 24 hours, photo captures of an endangered species with infrared camera in a natural reserve within the last month, or locations of particles in a refined space in a Brownian motion simulation. With need to handle specific spatial data comes spatial statistics, a branch of statistical tools that specifically analyse spatial point pattern and other types of spatial data (Harris et al., 2017; Munch et al., 2003; Wah et al., 2020).

One recent application of spatial statistics in data analysis research is analysing the pattern of cases of disease in epidemiology research. Infectious diseases have become one of the greatest challenges to humankind in the 20th and 21st century. Epidemiology and social research has demonstrated that infectious disease was one of the greatest extant challenge to humankind in terms of life quality, politics and economics, and both developing and developed worlds were ubiquitously challenged by infectious diseases (Fonkwo, 2008; Singh, Singh, 2020). Compared with other schools of statistical methods, one unique advantage of spatial statistics in epidemiology research is that: for many infectious diseases, the spatial information of identified cases of disease is highly important in determining the pattern of

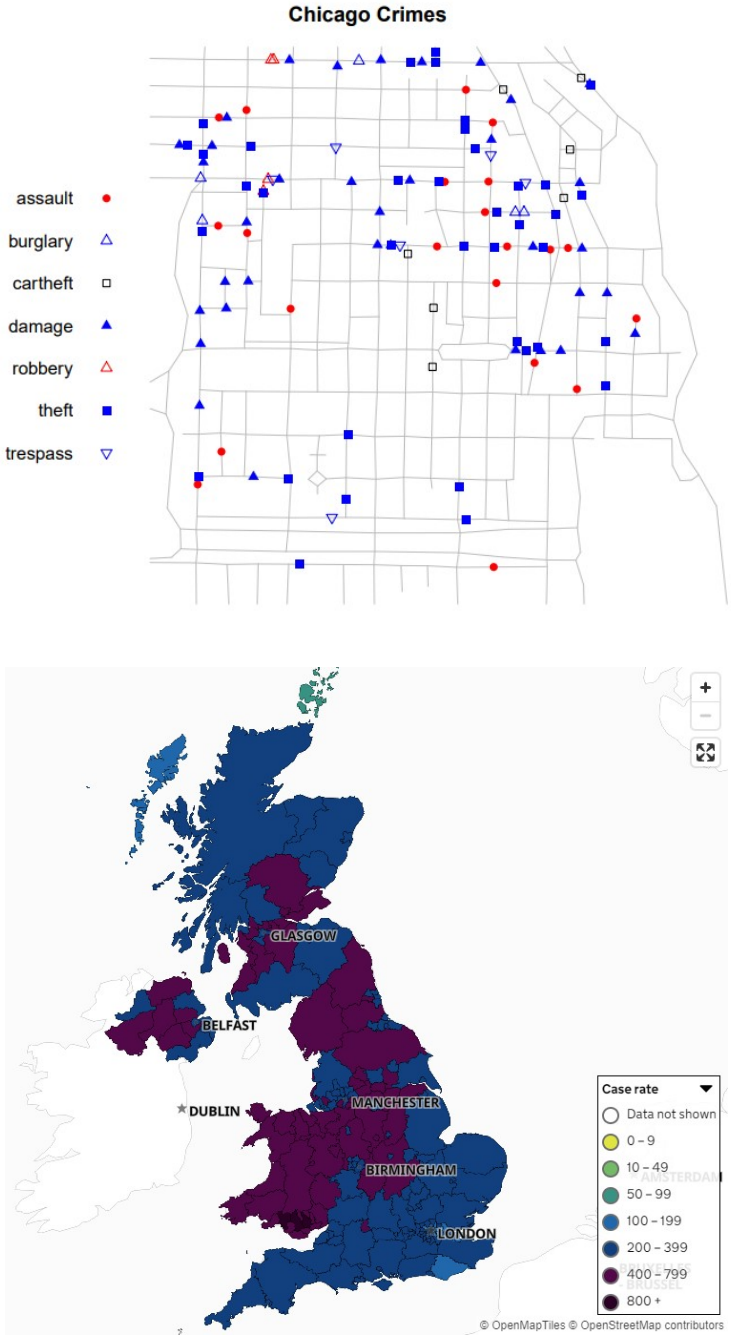


Figure 1: Visualisations of two different types of spatial data. Up, a spatial point pattern of locations of different crimes in a two-week period near University of Chicago. Down, spatial areal data of the per-region case rate of Covid-19 for the most recent 7-day period in the UK, reported in 3rd October, 2021. Scalebars were omitted for both data.

infection in subsequent times. With methods from spatial statistics can tools which analyse the pattern of infectious diseases be created; furthermore, models of spatial statistics based on existing observations can perform infectious disease surveillance and predict or alert potential pandemic outbreak. The use of spatial statistics in epidemiology research hence has high practical value. Currently a large amount of work on spatial statistics analysis has been done with numerous analysis methods and statistical models, conventional or bespoke, proposed (Harris et al., 2017; Moraga, 2020; Munch et al., 2003; Taylor et al., 2015; Wah et al., 2020).

This report about our spatial statistics analysis aims at introducing the concept of spatial statistics and some of the latest spatial statistics techniques to target audiences with background knowledge in statistics and data analysis, but little knowledge in spatial statistics. The report will start by introducing an example of spatial point pattern, followed by introduction to a series of current methods of exploratory data analysis specifically designed for spatial data, and finish by discussing spatial statistics model with respect to the data.

2 Overview of Spatial Point Pattern Data

2.1 Basic description of data

The main spatial point pattern data used in this analysis was collected as part of the AEGISS (Ascertainment and Enhancement of Gastrointestinal Infection Surveillance and Statistics) Project led by Prof. Peter Diggle at Lancaster University . A copy of the original dataset will be attached at the Acknowledgement section of the report. The AEGISS Project was established aiming to improve existing spatial statistics tools and infectious disease surveillance techniques (Diggle et al., 2005). The above dataset includes 10'572 "cases" of non-specific gastrointestinal infection (abbreviated as "NSGI" in this and all following sections). In this and all following sections, a case of NSGI, abbreviated as a "case", is defined as a report of non-specific gastrointestinal infection symptoms from an address in Hampshire (excluding the Isle of Wright) to the NHS Direct phone line, a 24-hour medical consultation service, between 1st January 2001 and 31st December 2003. Each case was verified by professionals checking reported symptoms against a list of common symptoms of non-specific gastrointestinal infections. In the context of spatial point pattern, a "case" in the Hampshire population might sometimes be called a "point" in the spatial point pattern, and the two words would be used interchangeably in this report. The original AEGISS data was downloaded from Prof. Peter Diggle's personal webpage in July, 2021,

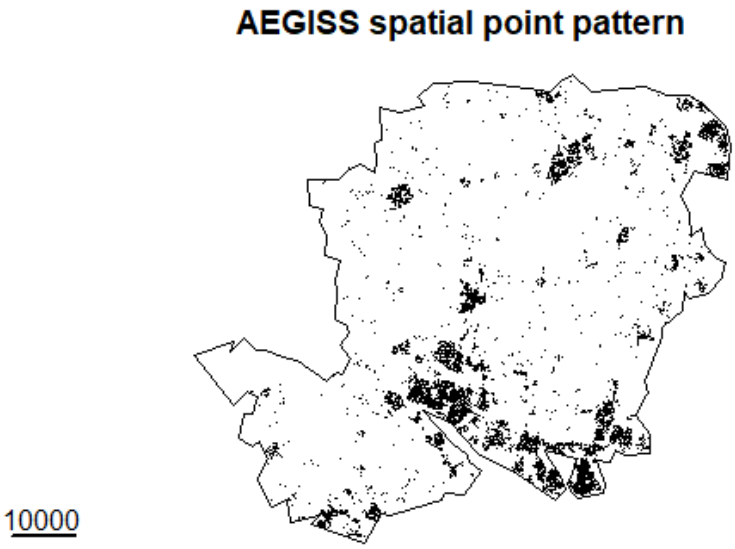
along with a shapefilewhich described the polygonal boundary of Hampshire. Each point in the dataset contained the following information: `id`, which acts as an internal reference for each individual case; `x` and `y`, a vector representing the reference location of the point (precise to the nearest metre); and `t`, a positive interget ranging from 1 to 1095 representing the calendar date at which the case was reported. The location vector `x` and `y` were taken as the coordinate of the centroid of the address at which the phone call to NHS Direct was made. `t = 1` indicates 1st of January, 2001 and `t = 1095` indicates 31st of December, 2003. All 10'572 data points were non-null.

Of all data points, 18 points were removed because locations of these points lied outside the boundary of Hampshire defined by the shapefile attached to the data. 441 points were considered for further removal because these points had the exact same combination of values of `x`, `y` and `t` with at least one other data point. If one point had the exact same values of `x`, `y` and `t` with the other point, it might suggest on date `t`, the same address with location (x, y) made two separate phone calls to NHS Direct about NSGI symptoms. It is unknown whether these two (or more) phone calls meant NSGI symptoms have appeared on two separate people living in the same address, or the same person had experienced NSGI symptoms twice. In the analysis I have chosen to eliminate all 441 duplicative points because a singular case in space and time is sufficient to indicate the presence of NSGI at the designated location and time. Hence, an updated definition of "case" in this analysis would be "at least one report within a full day from a specified address within Hampshire describing symptoms of NSGI to NHS Direct phoneline". There were further about 4000 points which had exact same values of `x` and `y` with other points, but not exact same values of `t`. These duplicative points might suggest that the same household has made multiple reports of NSGI to NHS Direct over the three-year surveillance period, and hence were kept in the dataset. A visualisation of the processed dataset could be seen in Fig. 2.

The main computational software for this analysis is `R`. The `R` package `spatstat` has been extensively used in the analysis.

2.2 Advantages and Disadvantages of AEGISS Data in Epidemiology Data Analysis

Compared with conventional disease data collected prior to AEGISS, the AEGISS dataset had advantage in accessibility and promptness. Patients needed no appointment to visit NHS Direct and to report their symptoms, hence the size of bias caused by report rate could be significantly reduced. The lack of need of laboratory test for a case to be verified (such as a laboratory test on patient faeces sample) also reduced



2a

2b

Figure 2: Up, the visualisation of the spatial point pattern of AEGISS dataset. Each black dot represents a case of NSGI as categorised in section 2.2. The black polygonal line represents the approximated border of Hampshire county in the UK. The short black line on the bottom-left was a scale bar of 10 kilometres. Right, a satellite map of Hampshire, UK, taken from Google Earth in 1st October, 2021 (ref website). The Isle of Wight was not sampled by the AEGISS Project.

the time lag between the infection event and the report of cases (Diggle et al., 2005; Diggle, 2013).

Consequently, a significant drawback of the AEGISS data was its low specificity because no test verifica-

tion was carried out. Therefore, Hence Diggle et. al. wrote in their paper that, significant aggregation of cases in the AEGISS data should better be treated as "anomaly" rather than "outbreak". Nonetheless, an "anomaly" in AEGISS data still had invaluable practical use. The abnormal density of reports of NSGI symptoms might not directly imply emergence of NSGI, but might be the indicator of other disease or public health problem (Diggle et al., 2005; Diggle, 2013). The other problem of AEGISS data is its vague definition of the sampling population. The dataset comes with no information on how well NHS Direct has covered Hampshire in 2001-2003, therefore, the people living in Hampshire did not directly equate to the people in Hampshire that can access NHS Direct. A home might not access NHS Direct because they were unaware of the existence of such service, or prefer the traditional appointment-based NHS consultation. It is therefore impossible or difficult to estimate the true size of population in Hampshire that can access NHS Direct and are willing to use such service to report NSGI symptoms. Further problem induced by the drawback of AEGISS data in exploratory data analysis will be discussed in section 3.2.

3 Exploratory Data Analysis of Spatial Point Pattern

3.1 Intensity in the context of spatial statistics

The first and most important analysis in the exploratory analysis of any spatial point pattern is the analysis on the pattern of intensity of the point process. A formal definition of intensity that considers spatial heterogeneity of intensity is as follows (Baddeley, 2016):

Definition 3.1 (Definition of spatial intensity)**.** *The spatial intensity, denoted as $\lambda(u)$, is a function of u, where u denotes the spacial location where the intensity is measured, such that in any given finite area A, the number of observed points from point pattern $\mathbf{X}$ that falls within the area A has the following expectation: $E(n(\mathbf{X} \cap A)) = \int_A \lambda(u)du$.*

An extended version of spatial intensity, called spatio-temporal intensity could therefore be given below as a reference to the AEGISS data:

Definition 3.2 (Definition of spatio-temporal intensity)**.** *The spatial intensity, denoted as $\lambda(u, t)$, is a function of u and t, where u and t respectively denote the spacial location and time where the intensity is measured, such that in any given finite area A and in any finite time interval T, the number of observed points from point pattern $\mathbf{X}$ that falls within the area A and interval T has the following expectation:*

$E(n(\mathbf{X} \cap A|T)) = \int_A \int_T \lambda(u,t)dudt.$

The analysis on the pattern of intensity has both theoretical and practical importance. Theoretically, the pattern of intensity helps spatial statisticians categorise the spatial point pattern and choose the correct type of spatial statistics model; practically, the quantity "intensity" could be directly translated to other concepts that have highlighted practical use. For example, the intensity of points in an epidemiological data (like AEGISS data) could be translated to the likelihood of observing the disease of interest in the area (Diggle et al., 2005).

A de-facto null hypothesis in the analysis of intensity would be that the intensity is homogeneous across space, meaning estimates of intensity should be similar between data sampled from different parts of the spatial point pattern. For AEGISS data with a temporal structure, the concept of homogeneity could be further expanded to denote similarity in estimates of intensity not only across space, but also across time. Contrasting with homogeneity means heterogeneity and heterogeneity within a spatial point pattern could have highlighted practical importance. For example, an area with significantly elevated intensity of cases in the AEGISS data could suggest potential outbreak of NSGI disease in the area (Diggle et al., 2005).

3.2 Adjustment to baseline risk level

In a full-scale spatial data analysis of data like the AEGISS data, the first stage before seeing any pattern of points or intensity is to adjust the data with the baseline incidence rate of NSGI. Such adjustment is necessary because any observed "heterogeneous" spatial pattern of cases might be the artefact of heterogeneous landscape: for example, if it is known that region A has twice the density of population of region$B$, it will become unsurprising to observe twice as many cases of NSGI in A than that in B (Diggle et al., 2005). This analysis could not adjust for the incidence rate of NSGI because no such suitable data was available. Data about Hampshire population between 2001 and 2003 were available and could possibly act as an substitute, if we assumed the baseline per-capita risk of NSGI was homogeneous across Hampshire; but most public data on Hampshire population did not have enough precision to be comparable to that of the AEGISS dataset. The precision of AEGISS dataset is to each single postcode-identifiable address and a comparable population data with similar precision might only be available through a population census. This analysis has therefore made an amateur assumption that both the population size and the incidence risk of NSGI in Hampshire was homogeneous. This assumption was clearly not well-held. However, suppose in subsequent analyses we have identified any unobserved effect

that could not be explained simply with spatial variation, we could test whether the unobserved effect was attributed to either population size or NSGI incidence risk by correlation analysis.

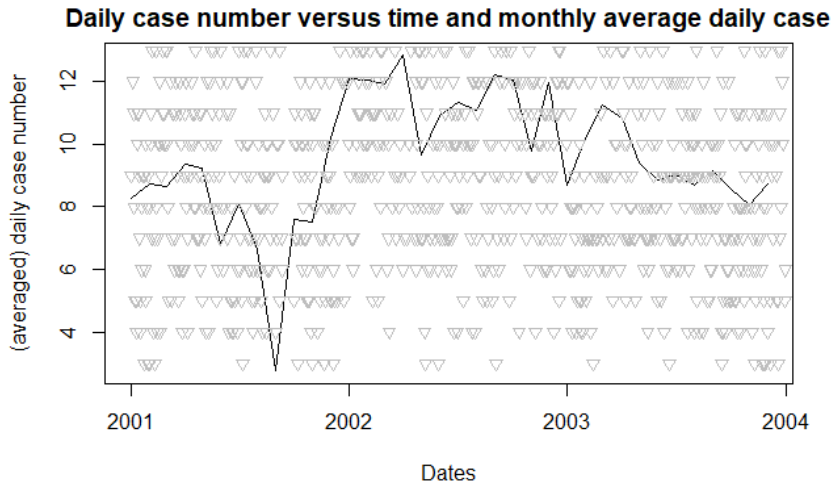
3.3 General temporal variation in the intensity

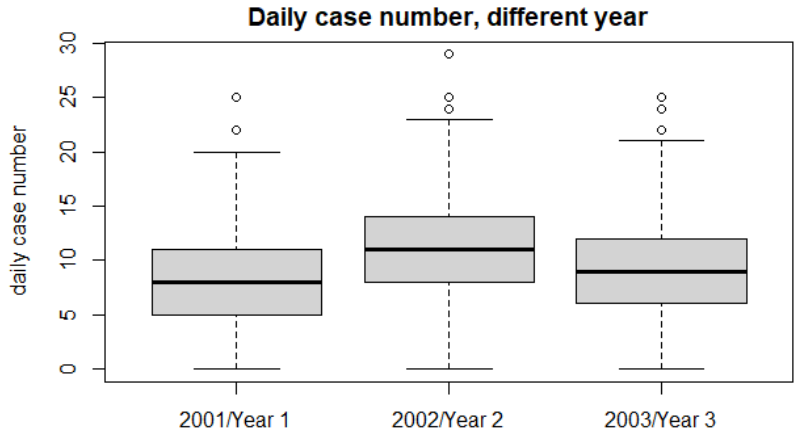
The temporal effect on patterns of cases is analysed in this and Diggle's analysis by observing temporal trends in daily case numbers. Diggle et al. has reported the following effects: 1) a time-of-year effect, where spring and autumns report higher case numbers than average; 2) a day-of-week effect, where weekend reported more average cases than weekdays; and 3) a general trend that daily case number increases with respect to time (Diggle et al., 2005; Diggle, 2013). To validate these earlier reports, I have again independently analysed time-of-year effect and day-of-week effect on daily case number with boxplots, analysis of variance and student t-test; we have analysed general temporal trend by plotting the trend of daily case number and monthly average daily case number against time. When analysing day-of-week effect we have manually divided the dataset into an integer number of weeks; since 1095 days (which is the duration of the sampling) did not equate to an integer number of weeks, we have only analysed data from the first 150 weeks, which corresponded to $t \in 1, 2, ..., 1050$.

Results of temporal trend analysis were displayed in Fig. 3. Overall, we did not obtain the result exactly as expected from Diggle et al's earlier study. There was sign of increase in daily case number in 2002 and 2003 compared with that in 2001 (Fig.3, Top); however, there was no seasonal pattern as described by Diggle et al., and the increase was not monotonic as the mean daily case number in 2003, compared with that in 2002, actually decreased (two-sided paired student t-test, assuming unequal variance: `mean(2002) = 11.49`; `mean(2003) = 9.29`; `p = 3.686e-11`). There were also no signs of increase in daily case number during weekends (two-sided, unpaired student t-test, assuming unequal variance: `mean(weekend) = 9.40`; `mean(weekday) = 9.65`; `p = 0.422`; Fig. 3, Middle). Daily case number distributions varied between different days of week (one-sided ANOVA: `p < 2e-16`), with Tuesday recording the lowest average case number. However, there seemed to be no clear, interpretable pattern (Fig. 3, Bottom).

3.4 Spatial variation in the intensity of spatial point pattern

Kernel smoothing estimator of intensity is a common non-parametric method to estimate spatial intensity in spatial statistical analysis. In the simplest term, kernel smoothing means that at each point u,





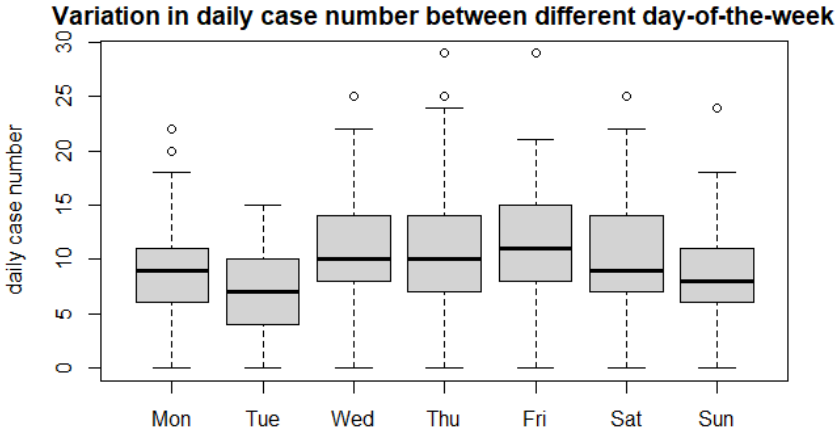


Figure 3: Patterns of daily case numbers across different scales of time. Top, grey dots represent daily case numbers, and solid black line represents trend of monthly averaged daily case number. Middle, boxplots of daily case numbers separated by year. Bottom, boxplots of daily case numbers separately by day-of-the-week, values on x-axis denote Monday, Tuesday, ..., Sunday respectively.

the estimate of intensity $\tilde{\lambda(u)}$ is calculated as the combination of not only u, but also points near u, hence $f(x)$ is "smooth" (Tan, 2015). The amount that each "nearby points" contribute to $\tilde{\lambda(u)}$ is then described by a probability density function known as the "kernel", usually denoted by $\kappa(u, x_i)$ where x_i denote any nearby points to u. Any kernel also comes with a bandwidth that, combining with the kernel, defines the concept of "nearby points" - what points in the spatial point pattern is considered "nearby" and how these points are weighted in the kernel smoothing estimator. For example, a disk kernel only considers points whose distance to u is within the bandwidth; a Gaussian kernel considers every point in the pattern irrespective of their distances to u, but each point is less weighted the further away they are from u - whose magnitude of weighting is determined by the σ, the common notation for bandwidth in Gaussian kernels (Baddeley, 2016; Tan, 2015). This analysis has used the kernel smoothing estimator of intensity with Diggle's correction and Gaussian kernel, which is defined as follows:
For any spatial location u,

$$\tilde{\lambda}^D(u) = \begin{cases} \sum_{i=1}^{n} \frac{1}{e(x_i)} \kappa(u - x_i) & u \in W \\ 0 & Otherwise \end{cases} \tag{1}$$

Where $\tilde{\lambda}^D(u)$ is the Diggle's corrected estimator of kernel smoothing intensity; $\kappa(u - x_i)$ is the Gaussian smoothing kernel; W denotes the space window where points were observed; x_i denotes the i_{th} point in the spatial point process $\mathbf{X}$ and $e(x_i) = \int_W \kappa(u - v)dv$ is a correction for edge effect(Baddeley, 2016; Diggle, 1985, 2013; Jones, 1993). In the case of AEGISS data, W would be defined as the space within the enclosed border of Hampshire. In this analysis the Diggle's corrected version of estimator has been selected because the estimator has smaller mean square error than other estimators and ensures the integral of $\lambda(u)$ over W is, as expected, always equal to the total number of observed points within the window (Baddeley, 2016). Gaussian kernel was a common choice in spatial statistical analysis and was hence selected since there is little prior information about the AEGISS dataset; Gaussian kernel was also the smoothing kernel used in (Diggle et al., 2005).

Calculations were performed with the `density.ppp()` function by `spatstat`. The algorithm of `density.ppp()` firstly estimated the intensity function $\lambda(u)$, with a pre-defined fixed smoothing bandwidth; the intensity was then realised at each pixel of a 248 $\times$ 248 pixellated field that approximated the space window W. The level of pixellation was chosen so that there should be generally more resulting pixels than the total number of points in the spatial point pattern. It was theoretically possible to construct a very fine pixellation by manually setting the dimension of each pixel to the minimum distance between points in

the spatial point pattern, so as to minimise the chance that two distinct points shared the same pixel. However, practically the resulting pixellation would involve 15.3 million pixels and became difficult to be computed. Additionally, the option `positive = "TRUE"` was called to force `density.ppp()` to always evaluate positive estimate.

The fixed smoothing bandwidth that must be supplied to `density.ppp()` could be calculated with several optimisation algorithms in `spatstat`. The choice of smoothing bandwidth is a trade-off between variance and bias: a smaller bandwidth will involve less bias but larger variance than a larger bandwidth. Most algorithms for smoothing bandwidth hence returned the bandwidth which optimised a defined target function. The likelihood cross-validation method maximised the point process likelihood cross-validation criterion (Loader, 1999):

$$LCV(h) = \sum_{i=1}^{n} log(\lambda_{-i}(x_i)) - \int_W \lambda(u)du \tag{2}$$

where LCV is the likelihood cross-validation criterion; $\lambda_{-i}(x_i)$ is the leave-one-out kernel smoothing estimate of intensity that was calculated using:

$$\lambda_{-i}(x_i) = \sum_{i \neq j}^{n} \frac{1}{e(x_j)} \kappa(x_i - x_j) \tag{3}$$

The other available method, the Diggle's cross-validated bandwidth selection, used a different optimisation target function that was described in (Diggle, 1985).

Diggle et al. has also used an adaptive bandwidth method in their estimations of spatial intensity. The adaptive bandwidth method used by Diggle et al. firstly performed an fixed bandwidth estimation with a pre-defined, initial fixed bandwidth h_0 and with equation 1, obtaining a "pilot estimate" of intensity $(\lambda_0(x_i))$ at each point x_i within the spatial point pattern $\mathbf{X}$. The adapted bandwidth h_i at each point x_i was then re-weighted as follows:

$$h_i = h_0\{\tilde{\lambda}_0 x_i / g\}^{-\frac{1}{2}} \tag{4}$$

where g is the geometric mean of the pilot intensity. Diggle et al. demonstrated with simulated heterogenous Poisson point process patterns that the adaptive bandwidth performed better than fixed bandwidth

in their analysis. Theoretically, adaptive bandwidth performed better than fixed bandwidth in analysis of clustered data because the adaptive bandwidth allowed more refined "smoothing" at areas with low relative intensity (Baddeley, 2016; Diggle et al., 2005).

The `spatstat` has implemented a different adaptive bandwidth algorithm which used the method developed by Abram, Hall and Marron (Abramson, 1982; Davies et al., 2018; Davies, Baddeley, 2018; Green et al., 1988; Hall, Marron, 1988). The adaptive bandwidth h_i was calculated instead with:

$$h_i = h_0 \cdot min(\tilde{\lambda_0}(x_i)^{-\frac{1}{2}}/g', trim) \tag{5}$$

where, g' is the geometric mean of $\tilde{\lambda_0}(x_i)^{-\frac{1}{2}}$, and $trim$ is a cut-off value to prevent the calculation of extremely small bandwidth values at lowly-densed part of the spatial point pattern.

In this analysis, all three bandwidth optimisation methods described above would be implemented to produce four separate sets of estimates of spatial intensity. The likelihood cross-validation method was implemented by `bw.ppl`; the Diggle's method was implemented by `bw.diggle`; and the Abramson, Hall, and Marron's adaptive bandwidth method was implemented by `bw.abram`. Of these, 1) and 2) were fixed bandwidth methods and 3) was the only adaptive bandwidth method. 3) was performed twice, where the initial fixed bandwidth h_0 was respectively recommended by `bw.ppl` and `bw.diggle`.

Results of estimations of spatial intensity was demonstrated in Fig4. As expected, there was strong signals of spatial heterogeneity like that demonstrated by the spatial point pattern plot (Fig2, Down). Particularly, all four estimations identified the following hotspots of cases at some big towns of Hampshire, namely Southampton, Eastleigh, Winchester, Basingstroke, Andover, Aldershot, Portsmouth and the outskirt area of Lymington. Portsmouth seemed to be a super-hotspot with the highest intensity of cases across the entire Hampshire. Considering that nearly all "hotspots" overlap with on-map major or minor human settlements, the result could not rule out the possibility that the observed spatial intensity pattern of NSGI was the artefact of heterogeneous distribution of population.

As for the effect of using different bandwidth selection criteria, the bandwidth optimised by `bw.diggle` has been much smaller than that by `bw.ppl` (Table 1). The effect of contrast in bandwidth size was hence clearly visible in Fig.4, as `bw.diggle` resulted in too refined and variable estimates of spatial intensity, whether the intensity was fixed or adaptive. Literatures suggested the use of `bw.diggle` for highly clustered dataset; however, given the variability of intensity estimates and the resulting plot, we have chosen to use adaptive bandwidth with starting bandwidth calculated by `bw.ppl` in the subsequent

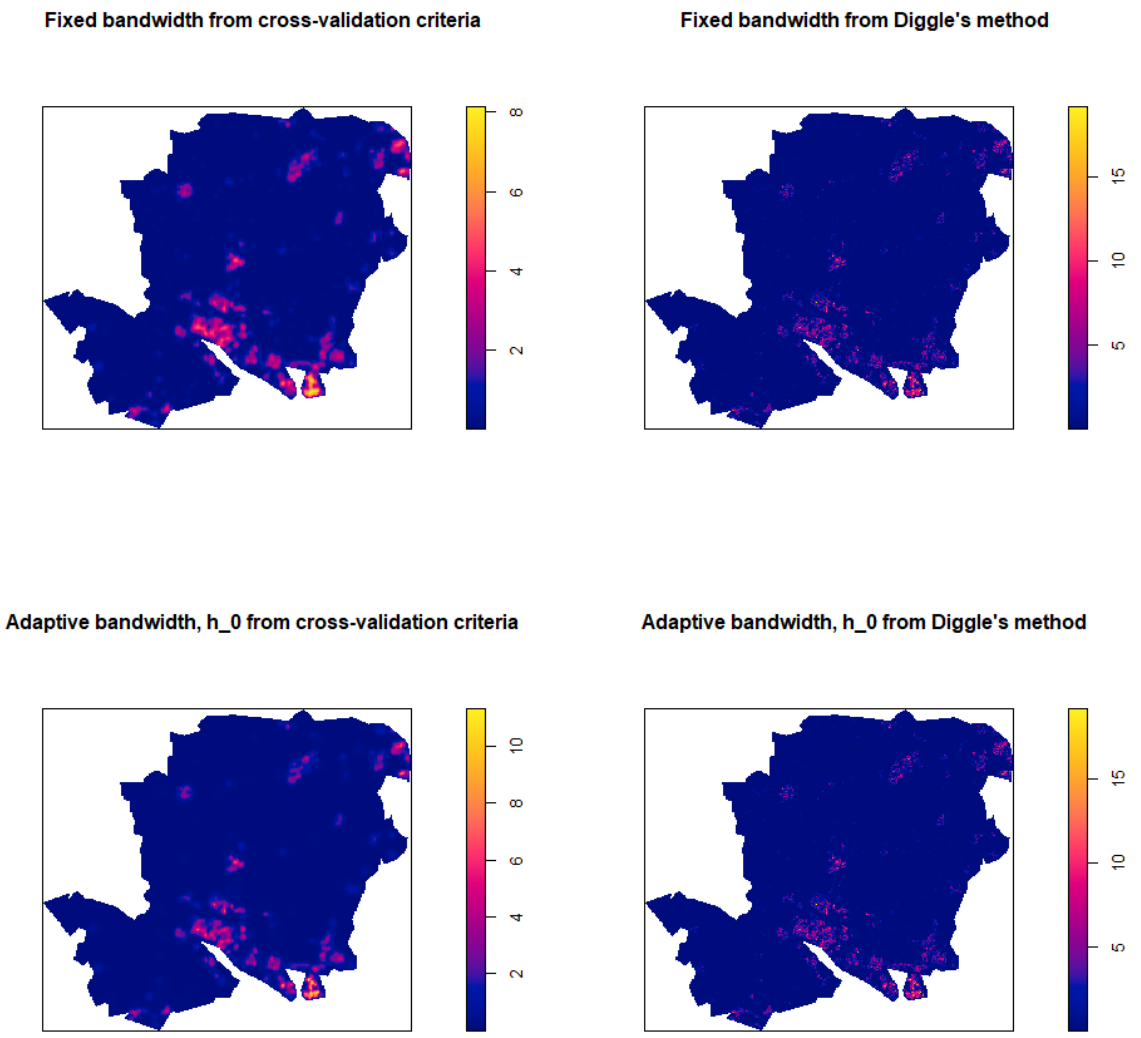


Figure 4: Kernel smoothing estimate of intensity of the AEGISS data across Hampshire, by four different methods of estimation. Topleft, fixed bandwidth optimised by cross-validation criterion; topright, fixed bandwidth optimised by Diggle's cross-validation; bottomleft, adaptive bandwidth with starting bandwidth optimised by cross-validation criterion; bottomright, adaptive bandwidth with starting bandwidth optimised by Diggle's cross-validation. Scalebars on the right of each plot have been manually inflated by a multiple of 100'000.

analysis. Estimated intensity produced by `bw.ppl` also better resembled the result in (Diggle et al., 2005). Additionally, there was no visible strong difference between the estimate from fixed bandwidth and adaptive bandwidth given that the starting bandwidth was the same.

Optimised bandwidth method	value of optimisation
cross-validation criterion	401.10
Diggle's method	10.57
adaptive bandwidth starting with cross-validation criterion	432.90 [219.69, 1023.68]
adaptive bandwidth starting with Diggle's method	8.90 [4.14, 15.48]

Table 1: Values of optimised bandwidths by different algorithms. For fixed bandwidth method, only the optimised value was present; for adaptive bandwidth method, the first number displayed the mean of adapted bandwidth and square bracket displayed the 2.5% and 97.5% quantile of all adapted bandwidth values.

3.5 Spatio-temporal variation in intensity homogeneity

Combining results from section 3.3 and 3.4, we were interested in whether there were interactions between the spatial and temporal components of variability in the pattern of points in the AEGISS data. For this analysis we will focus on the variation in spatial intensity pattern induced by yearly effects, that was discovered in section 3.2. To this, the AEGISS dataset was again separated by the year into three sub-groups; these three sub-groups were labelled "year-1 data", "year-2 data" and "year-3 data" based on the year of sampling in this and following sections. Each sub-group was then separately estimated for kernel smoothing intensity with both fixed and adaptive bandwidth. The starting bandwidth was the one optimised with `bw.ppl` on the combined dataset (that is, the dataset including all yearly data and the one that was used in section 3.3) and the same (starting) bandwidth for all three yearly datasets. This sacrificed optimisation level for each independent estimate but ensured that the level of bias and variance was consistent between all three sets of estimates (Baddeley, 2016).

The result of yearly estimates of intensity was demonstrated in Fig.5. The overall pattern of estimated intensity across Hampshire remained unchanged, but the overall magnitude of intensity has changed with respect to time, albeit the relationship was not monotonic. Year-1 estimate of intensity (Fig. **??**, top row) had significantly lower intensity at Portsmouth but comparable level of intensity at other hotspots than patterns of other years.

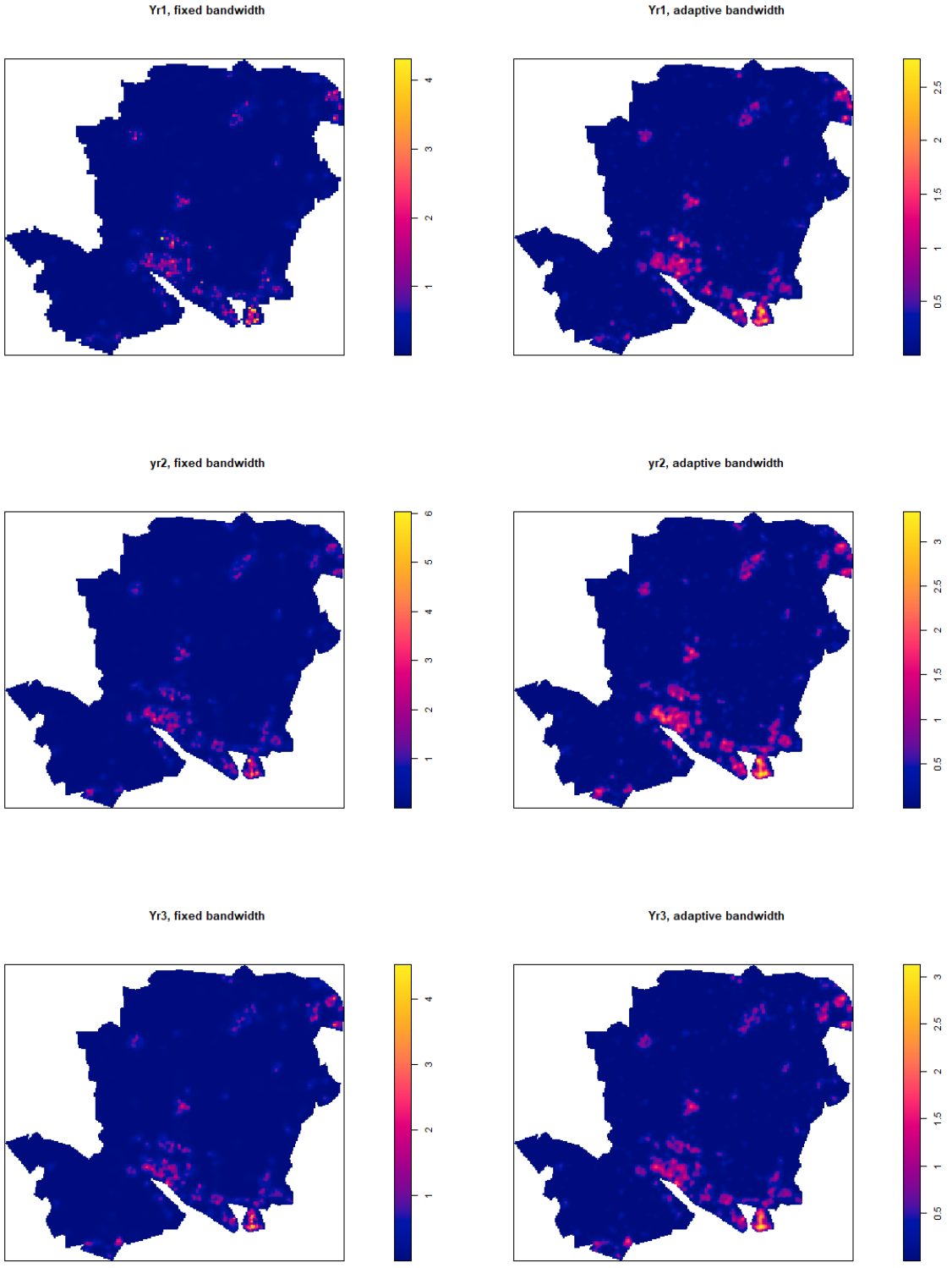


Figure 5: Kernel smoothing estimate of intensity for each of the three years from 2001 to 2003. Up row, 2001; middle row, 2002; bottom row, 2003. Left, fixed bandwidth; right, adaptive bandwidth. Intensities inflated by a factor of 100'000.

4 Spatial Statistical Modelling

4.1 Introduction to common spatial statistical models

Models of spatial statistics could be separated into two possible categories depending on whether the model assumes that dependency exists between separate points in the spatial point pattern. Models which assume no such dependency hence assumes that each point, at least in some ways, emerges independently from each other. This section would introduce several extant, commonly-used spatial statistics model in order of successive levels of understanding. Not every model mentioned in this section would be used in our analysis. Introductions of these models, however, would help readers understand

Poisson point process model is one of the most common models that assumes independence between points in the spatial point pattern. The entire Poisson point process model could be defined simply by two parameters: the intensity $\lambda(u)$, and the space window W. The Poisson point process model has the following assumptions (Baddeley, 2016):

Assumption 4.1. *The number of points within the spatial point process that are observed in any finite region $A \in W$, $n(\mathbf{X} \cap A)$, has a Poisson distribution;*

Assumption 4.2. *The number of points within the spatial point process that are observed in finite region A has the following expectation: $E(n(\mathbf{X} \cap A) = \int_A \lambda(u)du$;*

Assumption 4.3. *If A_1, A_2, ..., A_k are disjoint finite regions and for all i in $\{1, 2, ..., k\}$, $A_i \cap W$, then $(n(\mathbf{X} \cap A_1)$, $(n(\mathbf{X} \cap A_2)$, ..., $(n(\mathbf{X} \cap A_n)$ are independent random variables;*

Assumption 4.4. *Given that $n(\mathbf{X} \cap A) = N$, the N points are independently and identically distributed within A, with the same probability density:*

$$f(u) = \frac{\lambda(u)}{I} \tag{6}$$

where $I = \int_A \lambda(u)du$.

In homogeneous Poisson point process model, which assumes that intensity is homogeneous across W, $\lambda(u)$ is simply a constant λ. Similarly, the function in Assumption 4.4 could be reduced to $E(n(\mathbf{X} \cap A) = \lambda|A|$; the function in Assumption 4.4 could be reduced to $f(u) = \frac{1}{|A|}$, meaning points within A were uniformly distributed (Baddeley, 2016).

A key limitation of Poisson point process model is that $\lambda(u)$ is a function with respect to u. Following the basic property of function, the following implies that the same input of u always result in the exact same realisation of $\lambda(u)$ and the same predicted intensity at u. This limitation means any extrinsic random elements that might also interfere with the chance of developing NSGI symptoms have been completely discarded. As having mentioned in previous sections, the observed patterns of NSGI cases could be the artefact of many other unobserved elements, including but not limited to baseline population landscape and other environmental covariates of interest, such as local hygiene level, accessibility to health consultations, average age and report rate. These localised random effects could be incorporated with models such as the Cox model or the cluster model. The Cox model is, essentially, a Poisson process model where the intensity is random, instead of fixed, with respect to u. In terms of mathematics, Cox process replaced the term $\lambda(u)$ with $\Lambda(u)$, the latter being a stochastic process whose distribution is dependent on u. A common variant of Cox model is the log-Gaussian Cox process model, which additionally assumes that for any points $x_1, x_2, ..., x_n$ in the spatial point pattern $\mathbf{X}$, the driving intensity $\Lambda(x_1), \Lambda(x_2), ..., \Lambda(x_n)$ follows a multivariate log-Gaussian distribution (Baddeley, 2016; Cox, 1955; Møller et al., 1998).

The cluster model is a different type of spatial statistics model arising from Poisson process model and emphasises more the idea of "clustering". The first step of creating a realisation of cluster model still involves generating a Poisson spatial point pattern (that is, a spatial point pattern realised from the Poisson point process model). This spatial point pattern then becomes the "parental pattern" and each point in the pattern is used as the parental point to generate a "cluster" of offspring points; removal of each parental point from the pattern completes the generation of a cluster spatial point process. Depending on the method of generating clusters from parental points, there exists a variety of different cluster point process models with different assumptions. In this analysis we have used the modified Thomas cluster process model, which has the following assumptions (Baddeley, 2016):

Assumption 4.5. *Parental points form a Poisson point process;*

Assumption 4.6. *Different clusters are independent of each other;*

Assumption 4.7. *Different clusters have the same distribution given the same parental point;*

Assumption 4.8. *Given the parental point, offspring points within the same cluster are independently and identically distributed, and the probability density of offspring locations is isotropic Gaussian density;*

Assumption 4.9. *The number of offspring points generated by each cluster is a Poisson variable.*

The assumption 4.9 differs from the original assumption of Thomas model (Thomas, 1949) that the

number of points is a strictly-positive Poisson random variable; the algorithm used in `spatstat` has removed this constraint based on (Diggle, 1978). The resulting model will be referred as Thomas model in this and all following sections. The cluster model is probably a better tool in simulating real-world epidemiology problems, since the generation of offspring points from parents resembles the scenario where the disease was transmitted from one generation of patients to the next.

4.2 Modelling the intensity of AEGISS data

In the analysis by Diggle et al., (2005), the model proposed was an inhomogenous Poisson point process model where the spatially and temporally varying intensity $\lambda(u,t)$ is modelled as:

$$\lambda(u,t) = \lambda_0(u)\mu(t)R(u,t) \tag{7}$$

where $\lambda_0(u)$ represented purely spatial variability in intensity; $\mu(t)$ represented temporal variability in intensity; and $R(u,t)$ represented the belief of random elements that affect the intensity. Together, $\lambda_0(u)$ and $\mu(t)$ represented the fixed spatio-temporal effect on the variability of intensity.

One difference between our analysis and that of Diggle et al. is that, our analysis has a different conclusion in the temporal structure of the data. In the original literature, $\mu(t)$ was constructed by fitting a Poisson log-linear regression model and incorporated the original belief that there existed day-of-week and time-of-year effects. Since we only observed yearly effects, this analysis would take a different approach to model temporal variability in intensity in a different way. We have confirmed in section 3.3 and 3.5 that there existed a yearly effect on the variability of intensity. Therefore, it is possible to model $\lambda(u,t)$ as a step function as follows:

$$\Lambda(u,t) = \begin{cases} \Lambda_{year1}(u,t)R(u,t) & 1 \leq t \leq 365 \\ \Lambda_{year2}(u,t)R(u,t) & 366 \leq t \leq 730 \\ \Lambda_{year3}(u,t)R(u,t) & 731 \leq t \leq 1095 \\ 0 & Otherwise \end{cases} \tag{8}$$

where $\Lambda_{year1}(u,t)$, $\Lambda_{year2}(u,t)$ and $\Lambda_{year3}(u,t)$ could be estimated using yearly data.

Another contrast between our analysis and that of Diggle et al. is that the model of Diggle et al. has balanced for baseline incidence rate of NSGI, as mentioned in section **??**. Since it is impossible for us to balance for baseline incidence rate with the available data, a possible mitigation is to introduce an additional term $F(u,t)$ which represents the spatio-temporally localised fixed effect that was not captured in our model. Our final model is therefore:

$$\Lambda(u,t) = \begin{cases} \Lambda_{year1}(u,t)R(u,t)F(u,t) & 1 \leq t \leq 365 \\ \Lambda_{year2}(u,t)R(u,t)F(u,t) & 366 \leq t \leq 730 \\ \Lambda_{year3}(u,t)R(u,t)F(u,t) & 731 \leq t \leq 1095 \\ 0 & Otherwise \end{cases} \tag{9}$$

where $F(u,t)$ is a function of u and t.

In practice, the model would be fitted by fitting each of the yearly data into a spatial model that could take into account random effect $R(u,t)$. In this way, we were essentially modelling equation 8, not equation 9. The term $F(u,t)$ can therefore only be estimated by measuring the residuals of the resulting model from equation 8. This would be an imprecise estimate of $F(u,t)$, since the residuals from models based on 8 contained not only $F(u,t)$ but also residuals of the original model. All models would be fitted with the `kppm()` function(Guan, 2006; Guan et al., 2015; Jalilian et al., 2013; Tanaka et al., 2008; Waagepetersen, 2007).

4.3 Results

Fitted parameters of the spatial statistics model was shown in Table x. Here, for both log-Gaussian Cox process model and Thomas cluster model, `theta`, the intercept of the model, represented the logarithm of model-fitted expected intensity at any one point of the space window. The result of estimates of `theta` for both models partially contradicted with our earlier belief that spatial intensity in year-1, year-2 and year-3 data could be modelled with a step function that differed for each year: as shown in both table 2 and table 3, the 95% confidence interval of each estimate of `theta` for each year highly overlapped with that of estimates from other years in both models, suggesting that there was probably no significant difference in fitted spatial intensity between the three yearly models.

Residuals of both log-Gaussian Cox process model and Thomas cluster model show high correlation with estimated spatial intensity; areas of high residuals also overlapped with locations of some major

model: log-Gaussian Cox process		
`theta`	estimate	95% C.I.
year-1	-14.11	[-14.90, -13.31]
year-2	-13.71	[-14.50, -12.93]
year-3	-13.94	[-14.67, -13.20]

Table 2: Estimated values for parameter `theta` in the log-Gaussian Cox process model.

model: Thomas cluster process		
`theta`	estimate	95% C.I.
year-1	-14.11	[-14.84, -13.38]
year-2	-13.72	[-14.42, -13.01]
year-3	-13.94	[-14.61, -13.26]

Table 3: Estimated values for parameter `theta` in the Thomas cluster model.

towns in Hampshire that were identified earlier. Here, residuals were calculated by firstly using the `residuals.kppm()` function in spatstat (Baddeley, 2016); since the resulting residual values were too contrasting, residuals were then smoothed with the function `Smooth.msr()` which applied a Diggle kernel smoothing algorithm similar to that mentioned in section 3.4. As in Fig. 6, the most notable hotspot of residuals was located in Portsmouth, while other secondary-level residual hotspots were spotted in Southampton, Winchester, Andover, Bastingstoke and Aldershot. Locations of these hotspots of residuals collapsed with major towns in Hampshire (Fig. 2).

How do log-Gaussian Cox model and Thomas cluster model contrast with each other in performance? Since both models used different structrues and involved different parameters, it was difficult to directly compare the two models together. Instead, the log-Gaussian Cox process and the Thomas cluster model was compared with each other with simulation results. It should be noted that these simulation results could never approximate the real AEGISS data pattern, since we were fitting equation 8 and hence were not considering localised fixed effects; however, the simulated pattern could be a potential indicator of the performance of respective spatial statistics model. As shown in Fig. 7, Thomas cluster model has generated far too clustered spatial point pattern compared with the log-Gaussian Cox process model. The pattern generated by the log-Gaussian Cox process model therefore possibly better resembled the original AEGISS dataset as seen in Fig. 2, conditional on that the un-measured localised fixed effect did not significantly alter the simulated point pattern.

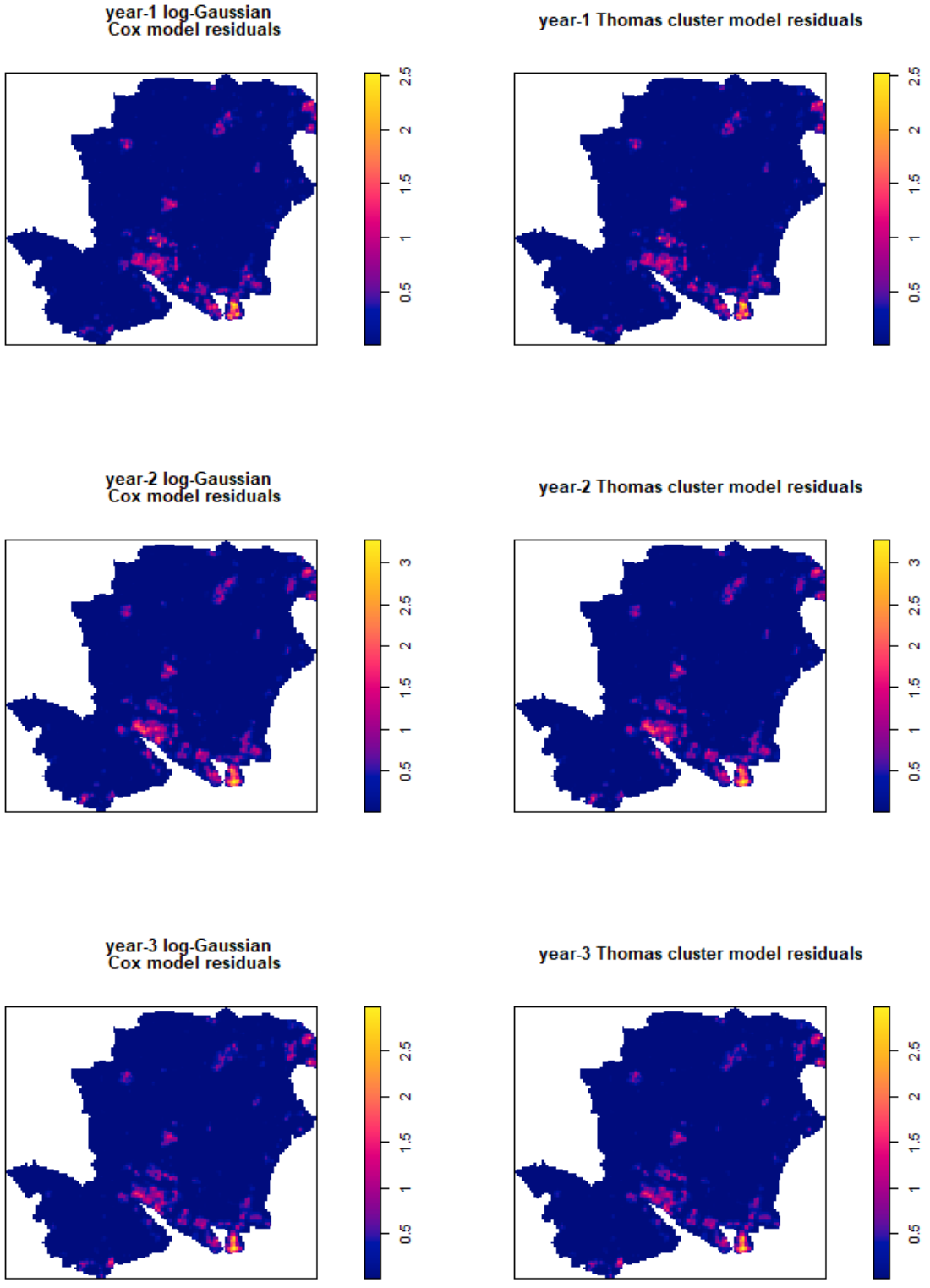


Figure 6: Smoothed residual plots of year-1, 2 and 3 models. Left, log-Gaussian Cox process model; Right, Thomas cluster model. Intensity inflated by a factor of 100'000.

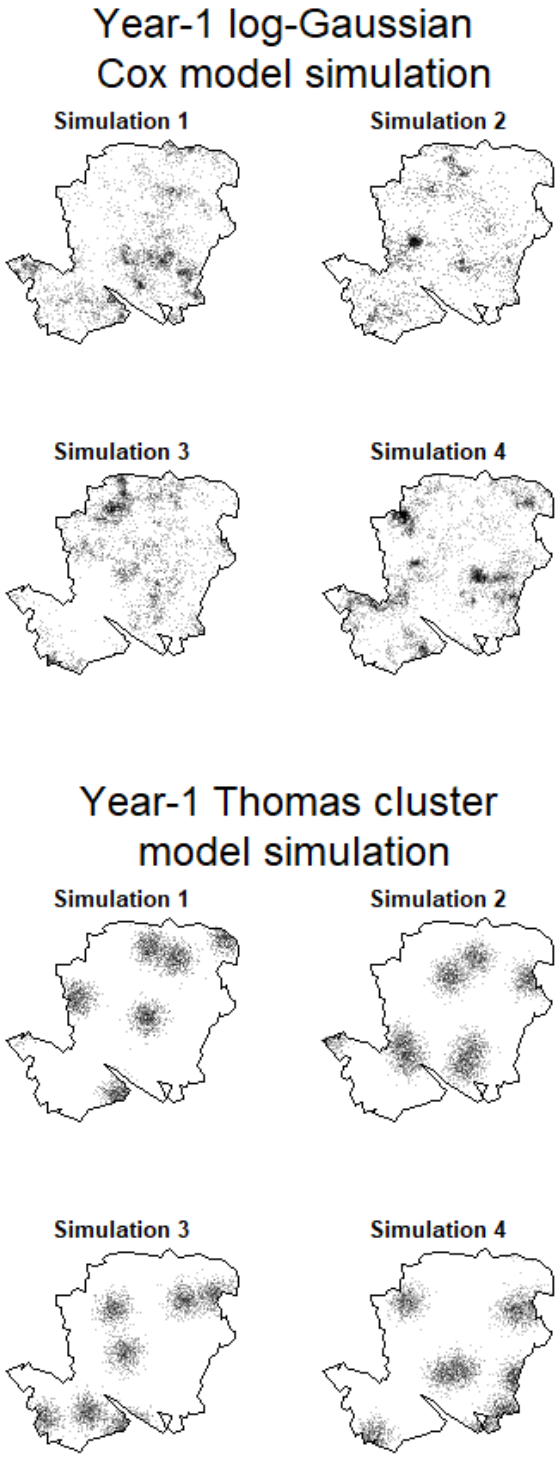


Figure 7: 4 independent realisations for both year-1 log-Gaussian Cox process model and year-1 Thomas process model. Only year-1 model simulations were displayed as examples.

5 Discussion

Our analysis has partially demonstrated the effectiveness of spatial statistics as a powerful tool in describing the underlying properties of spatial point pattern, or any other data with spatial structure. Even with the result of very simple, exploratory analysis result as shown in the section 3, the result could still prove great use conditional on that assumptions were held. For example, one finding in section 3 was that big towns such as Portsmouth had the highest incidence of NSGI in Hampshire county, which highlighted the need of distributing more NSGI-related medical resources in Portsmouth area; section 4.3 result on residuals possibly suggested that the underlying mechanism of the higher incidence of NSGI in Portsmouth could be caused by localised effect. Evidence of these localised effect therefore could have provided a new insight for researchers to explain the pattern of NSGI in Hampshire in the future, suppose better data become available.

Results of our analysis did not perfectly agree with that obtained by earlier researchers (Diggle et al., 2005). Parts of our results on the temporal variation of intensity contradicted with that obtained earlier (Diggle et al., 2005). Before questioning difference in these results, it should be noted that Diggle et al. had their analysis at a different focus compared with our analysis. In Diggle et al's analysis, all data have been adjusted with baseline incidence risk of NSGI before any statistical analysis. This adjust is probably the reason why our results did not agree with Diggle et al.'s. Diggle et al.'s performed such adjustment because the intention of AEGISS was to establish a practice-based surveillance tool whose main function was to alert "anomalous" peaks in cases of NSGI. Compared with Diggle et al.'s analysis, our analysis result did not highlight anomalies in incidences of NSGI; rather, our analysis is better interpreted as a general analysis on the distribution of NSGI cases in Hampshire. Conditional on the correct assumption and data, both Diggle et al's analysis and our analysis should be valuable in describing the trend of NSGI in Hampshire; however, results from these two analyses must be interpreted differently.

Due to limits in the available data, our analysis did not fully achieve the original target to fully describe the pattern of NSGI in Hampshire with spatio-temporal statistical model. This highlights the drawback of spatial statistics that it could be severely limited by the lack of very comprehensive data. For observatory analysis like ours, it is very difficult for one single statistician to obtain all necessary data that could potentially affect the result. unlike other formats of data, although spatial data were widely available, the format of spatial data could vary greatly depending on the accessibility of the statistician, and it might be difficult to combine different formats of spatial data. For example, the baseline population data of Hampshire was only available in spatial areal data format, with precision up to the level of 10

identified "major towns"; it was then difficult to apply the baseline population data in our analysis, since probably more than 50% of our cases were reported from these major towns. Time has also become a constraint to spatial statistical analysis, as temporal change in landscape plus the unavailability of required data in the past means statisticians could not predict the past with information available to the present. For example, for spatial data collected in the past at different administrative regions, the boundary of administrative regions might change over time; once the old definition of boundary became lost, it became difficult for later statisticians to uncover the original data without bias. Generally, it is easier for primary data analyst to analyse the spatial data since they, as the people who collect the data, have the best level of understanding to the data and can obtain the most up-to-date observations. For secondary data analyst like us, the difficulty of analysing spatial data could potentially be alleviated by future advancement in spatial statistics, better curation of spatial data and availability of better sampling techniques.

Finally, the problem caused by the lack of basic population landscape information could potentially be alleviated by the use of local composite likelihood method when fitting the statistical model. The local composite likelihood is similar to kernel smoothing estimation that a function similar to the kernel is used so that when the model is fitted with respect to a point, other points in the data were assigned weightings based on their distances to the fitted point. It was also mentioned that the local composite likelihood could reduce the problem in fitting inhomogeneous pattern that was induced by unknown background inhomogeneity. The local composite likelihood fitting could be performed with package spatstat.local in R; however the algorithm used was computationally demanding and hence could not be implemented in this analysis (Baddeley, 2017, 2016).

6 Acknowledgement

The code of the analysis is downloadable from this link, including all respective documents and data.

References

Abramson Ian S. On Bandwidth Variation in Kernel Estimates-A Square Root Law // The Annals of Statistics. 1982. 10, 4. 1217–1223. Publisher: Institute of Mathematical Statistics.

Spatial Point Patterns: Methodology and Applications with R. // . 2016. 830.

Baddeley Adrian. Local composite likelihood for spatial point processes // Spatial Statistics. Spatio-temporal Statistical Methods in Environmental and Biometrical Problems. XI 2017. 22. 261–295.

Cox D. R. Some Statistical Methods Connected with Series of Events // Journal of the Royal Statistical Society: Series B (Methodological). VII 1955. 17, 2. 129–157.

Davies Tilman M., Baddeley Adrian. Fast computation of spatially adaptive kernel estimates // Statistics and Computing. VII 2018. 28, 4. 937–956.

Davies Tilman M., Marshall Jonathan C., Hazelton Martin L. Tutorial on kernel estimation of continuous spatial and spatiotemporal relative risk // Statistics in Medicine. III 2018. 37, 7. 1191–1221.

Diggle Peter. A Kernel Method for Smoothing Point Process Data // Applied Statistics. 1985. 34, 2. 138.

Diggle Peter, Rowlingson Barry, Su Ting-li. Point process methodology for on-line spatio-temporal disease surveillance // Environmetrics. 2005. 16, 5. 423–434. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.712.

Diggle Peter J. On Parameter Estimation for Spatial Point Processes // Journal of the Royal Statistical Society: Series B (Methodological). 1978. 40, 2. 178–181. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1978.tb01660.x.

Diggle Peter J. Statistical Analysis of Spatial and Spatio-Temporal Point Patterns. VII 2013. 0.

Fonkwo Peter. Pricing infectious disease // EMBO reports. VII 2008. 9, S1. S13–S17. Publisher: John Wiley & Sons, Ltd.

Density Estimation for Statistics and Data Analysis. // . 1988.

Guan Yongtao. A Composite Likelihood Approach in Fitting Spatial Point Process Models // Journal of the American Statistical Association. 2006. 101, 476. 1502–1512. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Guan Yongtao, Jalilian Abdollah, Waagepetersen Rasmus. Quasi-likelihood for spatial point processes // Journal of the Royal Statistical Society: Series B (Statistical Methodology). VI 2015. 77, 3. 677–697.

Hall Peter, Marron J. S. Variable window width kernel estimates of probability densities // Probability Theory and Related Fields. XII 1988. 80, 1. 37–49.

Harris Nancy L, Goldman Elizabeth, Gabris Christopher, Nordling Jon, Minnemeyer Susan, Ansari Stephen, Lippmann Michael, Bennett Lauren, Raad Mansour, Hansen Matthew, Potapov Peter. Using

spatial statistics to identify emerging hot spots of forest loss // Environmental Research Letters. II 2017. 12, 2. 024012.

Jalilian Abdollah, Guan Yongtao, Waagepetersen Rasmus. Decomposition of Variance for Spatial Cox Processes // Scandinavian journal of statistics, theory and applications. III 2013. 40, 1. 119–137.

Jones M. C. Simple boundary correction for kernel density estimation // Statistics and Computing. IX 1993. 3, 3. 135–146.

Loader Clive. Local Regression and Likelihood. 1999.

Moraga Paula. Species Distribution Modeling using Spatial Point Processes: a Case Study of Sloth Occurrence in Costa Rica // The R Journal. 2020. 12, 2. 293.

Munch Zahn, Lill S, Booysen C, Zietsman H, Enarson DA, Beyers N. Tuberculosis transmission patterns in a high-incidence area: A spatial analysis // The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease. IV 2003. 7. 271–7.

Møller Jesper, Syversveen Anne Randi, Waagepetersen Rasmus Plenge. Log Gaussian Cox Processes // Scandinavian Journal of Statistics. 1998. 25, 3. 451–482. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9469.00115.

Singh Jaspreet, Singh Jagandeep. COVID-19 and Its Impact on Society. Rochester, NY, IV 2020.

Tan Henry. Kernel Smoothing Methods (Part 1). 2015.

Tanaka Ushio, Ogata Yosihiko, Stoyan Dietrich. Parameter Estimation and Model Selection for Neyman-Scott Point Processes // Biometrical Journal. 2008. 50, 1. 43–57. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.200610339.

Taylor Benjamin M., Davies Tilman M., Rowlingson Barry S., Diggle Peter J. Bayesian Inference and Data Augmentation Schemes for Spatial, Spatiotemporal and Multivariate Log-Gaussian Cox Processes in *R* // Journal of Statistical Software. 2015. 63, 7.

Thomas Marjorie. A Generalization of Poisson's Binomial Limit For use in Ecology // Biometrika. 1949. 36, 1/2. 18–25. Publisher: [Oxford University Press, Biometrika Trust].

Waagepetersen Rasmus Plenge. An Estimating Function Approach to Inference for Inhomogeneous Neyman-Scott Processes // Biometrics. 2007. 63, 1. 252–258. Publisher: [Wiley, International Biometric Society].

Wah Win, Ahern Susannah, Earnest Arul. A systematic review of Bayesian spatial–temporal models on cancer incidence and mortality // International Journal of Public Health. VI 2020. 65, 5. 673–682.