A SPATIAL ANALYSIS OF CASES OF NON-SPECIFIC GASTROINTESTINAL INFECTIONS IN SOUTHAMPTON.

32102717

Abstract

There are over 17 million cases of non-specific gastrointestinal infection every year in the UK, with incidence on the rise, and very low rates of reporting. We wish to investigate the relationship being the levels of deprivation in lower super output areas (LSOAs) in Southampton and the number of cases of non-specific GI infections. To do this we will utilise spatial models to take inferences about risk factors and areas of increased risk. We begin with exploratory analysis, followed by K-function analysis, intensity estimation, and fitting Poisson generalised additive models, and we finally we fit both non-spatial and spatial extra Poisson variation models. Our analysis suggest that the Index of Multiple Deprivation is not a good explanatory variable for the risk of GI infections, but a subset of its domains may be better. There is, however, still a lot of unexplained spatial variation present in the models, which suggests there are other variables which we have not considered which may play a big role in the risk associated with GI infections. Finally, we suggest some ways in which the models can be improved, including including more variables that could be risk factors, using individual level data, or adding a temporal component as well as spatial component.

---

1. Introduction

Non-specific gastrointestinal infection encompasses any disease that infects the gastrointestinal tract, which includes the oesophagus all the way to rectum, and accessory digestive organs such as the liver, gall bladder and pancreas[9]. It is estimated that 25% of the UK population suffer from some form of infectious intestinal disease every year[4] (also known as non-specific gastrointestinal disease[4] or gastroenteritis[6]). The incidence of infectious intestinal disease is on the rise, with a 43% increase in 2008-09 when compared with 1993-96[4]. It was estimated that for every case of IID presented to national surveillance, there were an additional 147 cases in the community[4]. Of these more than 17 million annual cases, only 2% visit their GP, which is a decrease of 50% compared to 1993-96[4]. Today, this may be partially explained by elements like NHS Inform Scotland suggesting that individuals do not visit their GP, as the infections are easily spread, but instead ring the NHS 24 111 service[11]. A comprable system, NHS direct, now defunct, was in effect in 2008-09, but only accounted for a very small proportion of case detection[4]. This makes incidence of IIDs and their effects on the community very difficult to estimate, even though reporting of IID to national statistics by GPs has improved[4]. This represents a large impact on the UK, with around 50% of cases reporting IID-related absences, leading to 11 million missed work days and 8 million missed school days[4].

It is well known that for many chronic diseases, and some infectious diseases, socio-economic circumstances affect risk in individuals[5;13;14]. In fact, poor socio-economic standing can affect all aspects of health throughout life, for example, in Scotland, individuals born in the most affluent areas of Glasgow live 10 years longer than those born in the most impoverished areas[15]. For some infectious diseases, there exists evidence that incidence varies by socio-economic standing, such as tuberculosis or HIV[12]. For non-specific GI infections, however, the relationship is not well understood[12]. Some studies suggest that it is in fact higher socio-economic status (SES) that are related with increased burden[7;10], but also show that these relationships may cease to exist once one adjusts for recent foreign travel[7], or that they aren't consistent across all pathogens[10]. Thus discerning the relationship between all infections that are encapsulated in "non-specific GI infection" and SES is not a clear task.

Relationships between SES and prevalence of disease are visible whether one measures at the individual level or deprivation by area, by utilising a deprivation index, such as the Index of

Multiple Deprivation (IMD) [15]. This is good as data is usually easier to collect and more available at the area level. The differences in exposures for different levels of deprivation could have many explanations. Many studies suggest that different pathogens are related with increased risk in different SES groups, for instance some found that increased SES increases the burden of infections caused by Campylobacter, E. coli, and Salmonella, but decreases the burden of Listeria [7;10]. There are over 250 possible infectious and non-infectious agents that can contaminate food [10]. The differences, however, could arise from not adjusting for the correct variables, such as recent foreign travel [7] as previously mentioned. These variables, however, could to some degree suggest differences in SES. The differences are supported by the fact that the different SES groups are asscoaited with the different risk factors for the disease [10]. Another possible explanation is that different socio-economic groups have different diets and food preparation hygiene levels, which could affect their exposure [10]. For instance, there is evidence to suggest that increased SES increases the likelihood that one will consume undercooked/raw food, such as raw oysters or rare beef, whereas lower SES is associated with insufficiently cooled refrigerators [10]. Neither, however, is yet to be associated with increased risk [10]. There are some variables that have been found to be associated with IID (at least for some organisms), such as income, higher educational level, home ownership greater than 50% at the community level, and semi-routine occupations [10].

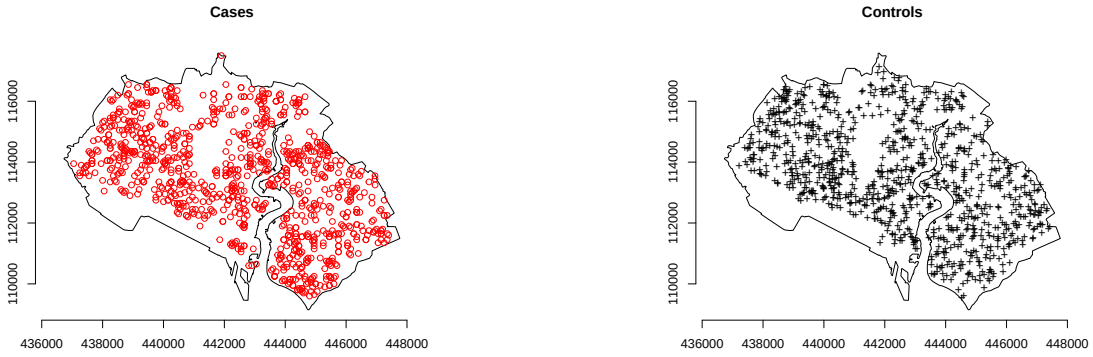
The differences, though, could also be more down to the metric and tools used to measure differences in population and incidence. For instance, it was found that different SES metrics were not consistent in their results, however, deprivation indexes were [10]. There are still issues with deprevation indices though. For instance, the Index of Multiple Deprivation looks at 37 separate indicators in each Lower Super Output Area (LSOA), and combines them by weighting each element [8]. Within each area, however, there may be poor areas and richer areas, or different residential patterns associated with different ethnicitities [5]. It may also miss important factors such as the feeling of belonging and community [5]. Another important factor could be reporting bias [10], it was found that, for IID, increased levels of deprivation are associated with the rate of presentation to primary care [7]. Also that hospitalisations association with SES was one of the strongest for GI infections, when compared to other conditions [1]. Even once these relationships are investigated, and the associations determined, some suggest that only a limited set of interventions are utilised, and it would be better to choose the intervention that best fits the scenario [13], perhaps based on organism.

In this report we wish to look at spatially aggregated data from the Ascertainment and Enhancement of Gastrointestinal Infection Surveillance and Statistics (AEGISS) project [3] in Southampton collected between 1 January 2001 and 31 December 2002. The data contain information on 1000 cases of non-specific gastrointestinal infection and their locations, as well as information used in the Index of Multiple Deprivation (IMD) and the IMD score for each lower super output area (LSOA) in Southampton. The data was collected using the NHS Direct telephone clinical advice service, now defunct. We also have a set of controls simulated based on population density in the area. We will begin in §2 by considering simple exploratory data analysis, plotting the cases and controls in space, and looking at differences in the distribution of cases for different variables after overlay operations. In §3 we utilise K-function analysis to try and identify whether there is clustering in the cases and controls, and compare the two. In §4 we estimate the intensity of the process across the plane and use graphical aids to interpret it, and compare the results for different methods of bandwidth selection. Following that, in §5 we use Poisson Generalised Additive Models fit with a subset of variables to identify areas that have an increased expected number of cases. In particular, we focus on whether the IMD or a subset of its components are a better indicator of case count per area. Our analysis culminates in §6 by fitting Bayesian non-spatial and spatial extra Poisson variation models, taking into account associations with neighbouring LSOAs, and we investigate the random effects for evidence of unexplained differences between the LSOAs. Finally we discuss our findings and take inferences in §7, as well as considering some ways of improving the analysis in the future.

2. Exploratory Analysis

The data is available to us at two levels. The first is that of the individual level, which tells us which individuals are cases and where they are on the plain. The second is an aggregated level, which contains the deprivation characteristics of each of the 147 LSOAs in Southampton, their borders and centroid locations, and some additional information like population size.

At the individual level there are 1000 cases and 1000 simulated controls. They do not appear to be distributed uniformly throughout the plain, as can be seen in Figure 3, however, this is best left to K-function and intensity estimation. The study covers an area of approximately 4,984,000 square units, which is home to approximately 212,000 people. Each LSOA has a population between 1,100 and 2,400, with a median around 1500, on average about half of which are male. The IMD scores of the LSOAs range from around 7 to 57, with a median of 21, but a mean of 24. Every LSOA had atleast one case, with the median being around 7, and the most in one area being 14.



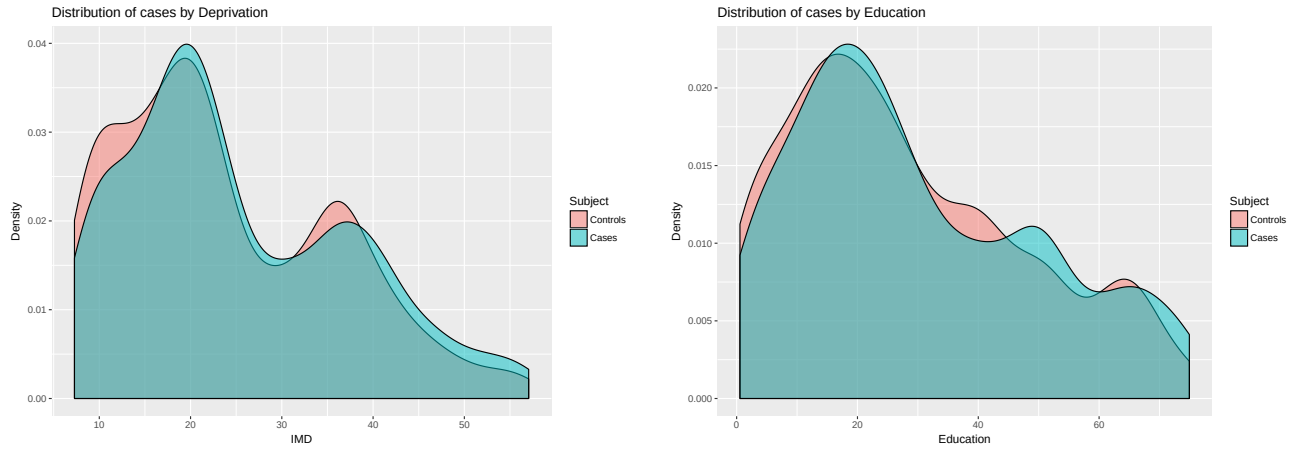
(A) (B)

Figure 1. (A) A visual representation of the locations of the cases throughout Southampton. (B) A visual representation of the locations of the controls throughout Southampton.

We can utilise overlay operations to combine the two levels of data, and investigate whether there are any differences in the distribution of cases and controls for different variables. From Figure 2 we can see that, in general, there are more cases for higher levels of deprivation. Figure 2a shows that there seem to be more cases than controls in more deprived areas, and Figure 2b suggests more cases occur when the score of the education domain is higher (so the education level is lower). Other variables demonstrated similar relationships. All figures, however, also have a spike to the left of the graphs, suggesting there may also be some relationships with certain lower levels of deprivation.

3. K-function Analysis

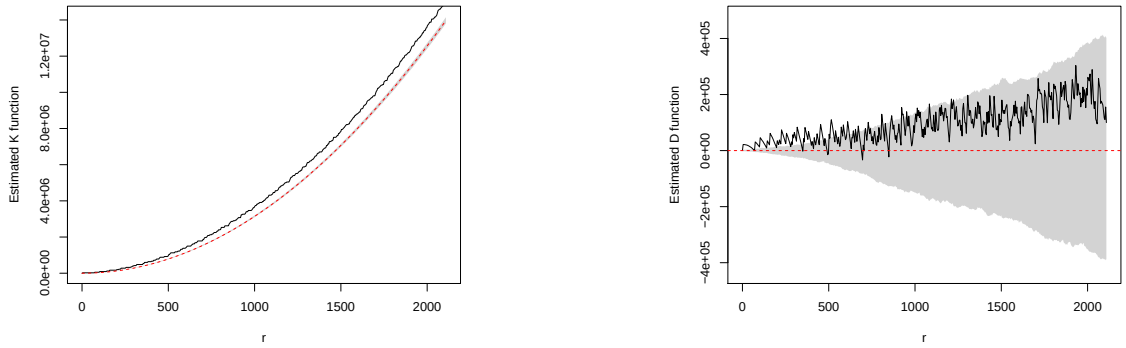
It is important to know whether the cases are clustered in space. To do this we can use K-function analysis, which for each cases measures the expected number of cases surrounding it, and is calculated with a range of different radii. We will utilise isotropic edge-correction methods. If the data comes from a homogeneous Poisson process then we would expect that $K_{jj}(s) = \pi s^2$, where j represents that we are looking at cases surrounding cases, and s is the radius of the circle surrounding the point. We can see from Figure 3a that the estimated K-function for the cases deviates greatly from the confidence interval, suggesting the cases are clustered in space. We would, however, expect the cases to be clustered in space, if the population they come from is also clustered, for instance, how human populations cluster around urban areas. Thus, under the null hypothesis, we can say there is no spatial clustering if the cases and controls are independent samples from the same underlying population at risk, such



(A) (B)

FIGURE 2. (A) Density plots for the distribution of the cases by Index of Multiple Deprivation score. Larger scores represent greater levels of deprivation. (B) Density plots for the distribution of the cases by the lack of attainment and skills in the area.

that $K_{ii}(s) = K_{jj}(s)$, where i is controls and j is cases. Hence, we can consider the difference between the two, $D(s) = K_{ii}(s) - K_{jj}(s)$, and look at whether it is significantly different from 0. We can see from Figure 3b that, for the most part, $D(s)$ is within the confidence bounds, meaning that there is probably no spacial clustering in the case process.



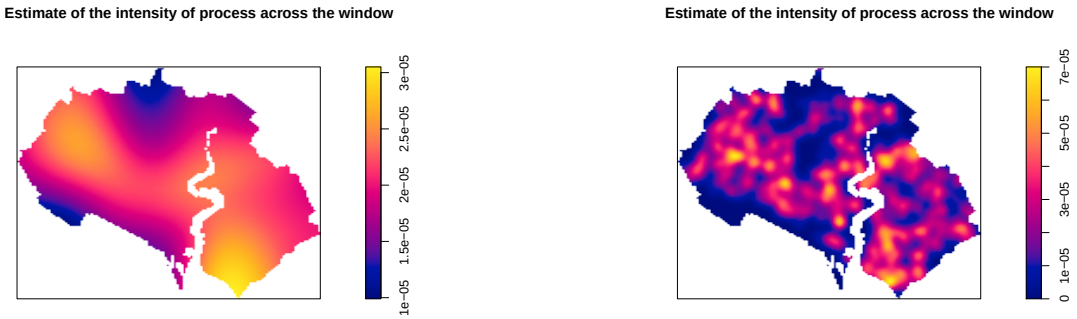
(A) (B)

FIGURE 3. (A) A plot of the K-function for a range of radii checking for clustering among the cases, with a 95% confidence bound. (B) A plot of the difference between the K-functions for the cases and controls at a range of radii, checking for spatial clustering in the case process, with a 95% confidence bound.

4. Intensity Estimation

It can also be useful to estimate the intensity of the process across the plain. This can help identify areas where more cases are occurring, which could give an insight into the risk factors associated with the disease. Intensity estimation (or density ratios) uses kernel smoothing methods to estimate the spatial variation in risk. The level of smoothing depends on a parameter, h, which dictates the bandwidth. Figure 4a suggests two areas where there is, relatively, a much greater risk of being a case as opposed to a control, one in the south-east and one in the north-west. For this estimate the value of h was chosen by default using a simple rule of thumb,

and had a value of $h = 1054.44$. We can, however, choose the bandwidth using cross-validation methods, to use a bandwidth which best matches the data. One option is to use a method proposed by Diggle, with a calculated bandwidth of $h = 103.17$. It does not suggest any strong evidence for spatial variation in risk (plot omitted). An alternate cross validation method is used in Figure 4b, with a calculated bandwidth of $h = 169.95$. This plot also suggests some spatial variation in risk, with areas in the south-east and north-west again, and also maybe in the north-east.



(A) (B)

FIGURE 4. (A) An estimate of the intensity of the process using a rule of thumb bandwidth. (B) An estimate of the intensity of the process using an alternative cross validation method to calculate the bandwidth.

5. Generalised Addative Models

While intensity estimation is useful to try and identify areas of increased risk, it does little more. An alternative method we can use to estimate spatial variation in risk is the Generalised Additive Model (GAM). GAMs are an extension to the traditional generalised linear model that allows the relationship between the response variable and covariates to be estimated as part of the model-fitting procedure. The general form of the model for Poisson regression is such that:

$$\ln(\mu_i) = u(x_i)'\beta + g(x_i), \tag{1}$$

where μ_i is the expected number of cases in the i^{th} LSOA, $u(x_i)$ is a vector of covariates for the i^{th} LSOA, β is a vector coefficients, and $g(x_i)$ is a function that models smooth residual spatial variation.

We begin by fitting a Poisson GAM to model the number of cases in each LSOA. The data contains covariates for the score of the IMD and the respective scores of the domains that contribute to it, as well as some basic population counts. We begin by fitting a model with only the spatial variation term, offset by population size in each LSOA, of the form:

$$\ln(\mu_i) = \log(pop_i) + \beta_0 + s(xcoord_i, ycoord_i), \tag{2}$$

where pop_i is the population size in the i^{th} LSOA, β_0 is the coefficient estimate of the intercept, and $xcoord_i$ and $ycoord_i$ are the coordinates of the centroid of the i^{th} LSOA. Figure 5a shows the fitted residual surface for this model, we can see that there is a lot of unexplained spatial variation. Next we can try adding the Index of Multiple Deprivation score of each LSOA as a covariate, and see if it can explain some of the spatial variation. The model has the form:

$$\ln(\mu_i) = \log(pop_i) + \beta_0 + \beta_1 U_{\text{IMD},i} + s(xcoord_i, ycoord_i), \tag{3}$$

where pop_i is the population size in the i^{th} LSOA, β_0 is the coefficient estimate of the intercept, β_1 is the coefficient estimate of the IMD score for each LSOA, $U_{\text{IMD},i}$, and $xcoord_i$ and $ycoord_i$

are the coordinates of the centroid of the i^{th} LSOA. The coefficient for IMD is not statistically significant at the 5% level (p-value = 0.086), and we can see from Figure 5b that it does not explain much of the excess spatial variation when compared with Figure 5a.

Instead, we can investigate whether breaking the IMD into its components and allowing each one to have its own coefficient allows us to explain more of the spatial variation in risk. We fit a model with the form:

$$\begin{aligned}\ln(\mu_i) = \log(pop_i) + \beta_0 + \beta_1 U_{\text{Income},i} + \beta_2 U_{\text{Employment},i} + \beta_3 U_{\text{Health},i} + \beta_4 U_{\text{Education},i} \\ + \beta_5 U_{\text{Barriers},i} + \beta_6 U_{\text{Crime},i} + \beta_7 U_{\text{Enivironment},i} + s(xcoord_i, ycoord_i),\end{aligned}$$

where pop_i is the population size in the i^{th} LSOA, the β's are the coefficient estimates, the $U_{\bullet,i}$ are the covariate values for the i^{th} LSOA, and $xcoord_i$ and $ycoord_i$ are the coordinates of the centroid of the i^{th} LSOA. In this model only employment is statistically significant at the 5% level (p-value = 0.004), but from Figure 5c we can see that there is still a lot of unexplained spatial variation. We always strive for the most parsimonious model, so we will utilise backwards elimination to select a subset of the variables. This results in a model of the following form:

$$\ln(\mu_i) = \log(pop_i) + \beta_0 + \beta_1 U_{\text{Employment},i} + \beta_2 U_{\text{Education},i} + s(xcoord_i, ycoord_i), \tag{4}$$

where pop_i is the population size in the i^{th} LSOA, the β's are the coefficient estimates, the $U_{\bullet,i}$ are the covariate values for the i^{th} LSOA, and $xcoord_i$ and $ycoord_i$ are the coordinates of the centroid of the i^{th} LSOA. In this model both Employment and Education are statistically significant at the 5% level (p-values = 0.0005 and 0.035 respectively), but from Figure 5d we can see that there is still a lot of unexplained spatial variation. Using the AIC we can compare the models, and we see that the final model fits the data the best, but the plot of the fitted residual surface shows there is a lot of unexplained spatial variation, meaning there is most likely other variables that we have not adjusted for that contribute heavily to the risk.

6. Bayesian Models

There is evidence to suggest that there is some unexplained spatial variation in the model which we are not accounting for. For instance, we can attempt to fit a generalised linear model to the count data, however, when we do we find, using a χ^2 goodness of fit test, that the nominal standard errors from the Poisson regression model are too small. The test statistic is approximately $X^2 = 173.9$ which is around 20% larger than $n-p = 144$. Thus, there is evidence to suggest that the data is overdispersed. Now that we know this, a better fitting model would be an extra Poisson variation model, which takes the general form:

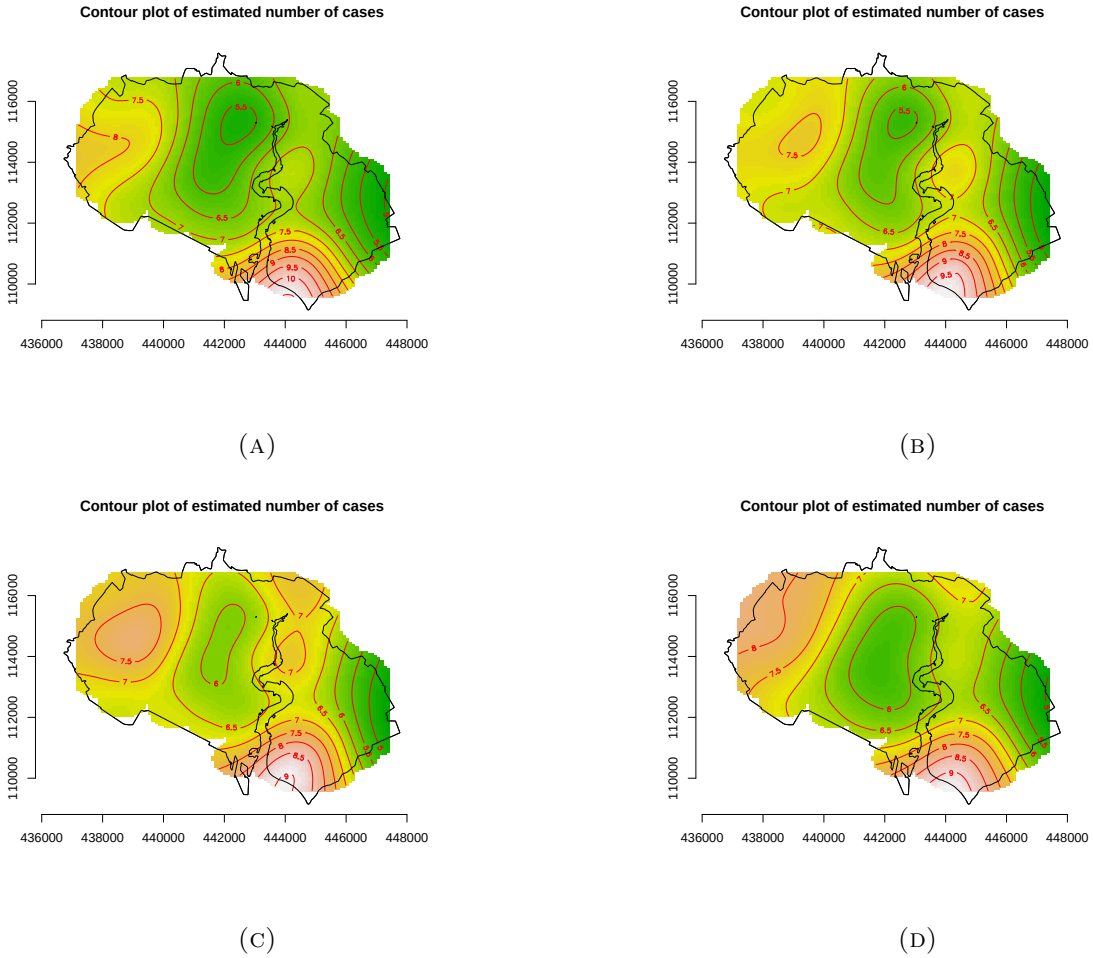
$$\log(R_i) = \alpha + \beta x_i + U_i + S_i, \tag{5}$$

where α is the intercept, β are the coefficients to the covariate values x_i, U_i are mutually independent $N(0, \nu^2)$ random effects without any spatial structure, the S_i are spatially correlated random effects that follow a discrete spatial variation model in which two LSOAs are neighbours if and only if they share a common boundary. The full conditional distributions of LSOA i depend only on its neighbours, and $S_i|\text{neighbours} \sim N(m_i, v_i)$ where m_i is the mean of the S_j from LSOAs j which are neighbours to LSOA i, and $v_i = \sigma^2/n_i$, where n_i is the number of neighbours of LSOA i.

We begin by fitting a non-spatial extra Poisson variation model, our chosen model has the form:

$$\log(R_i) = \log(pop_i) + \alpha + \beta_1 x_{\text{Employment},i} + \beta_2 x_{\text{Education},i} + U_i, \tag{6}$$

where α is the intercept, the β's are the coefficients of the covariates $x_{\bullet,i}$, and U_i are mutually independent random effects without any spatial structure. These models are fit using MCMC methods, so before we can take any inference from them we need to check that they have converged and that they are mixing well. From trace and density plots (omitted), all variables appear to be mixing well, and from Figure 6a we can see that, at lag 5, all estimated parameters have reasonable ACF values, though the variance parameter governing the variance of the random effects, τ^2, may be a little high, but still within reason given all the others. For this model, Education was not significant, as its confidence interval crossed 0, but Employment



(A) (B)

(C) (D)

FIGURE 5. (A) Fitted residual surface for GAM with only spatial component. (B) Fitted residual surface for GAM with IMD as a covariate. (C) Fitted residual surface for GAM with IMD components as individual covariates. (D) Fitted residual surface for GAM with subset of IMD components as individual covariates.

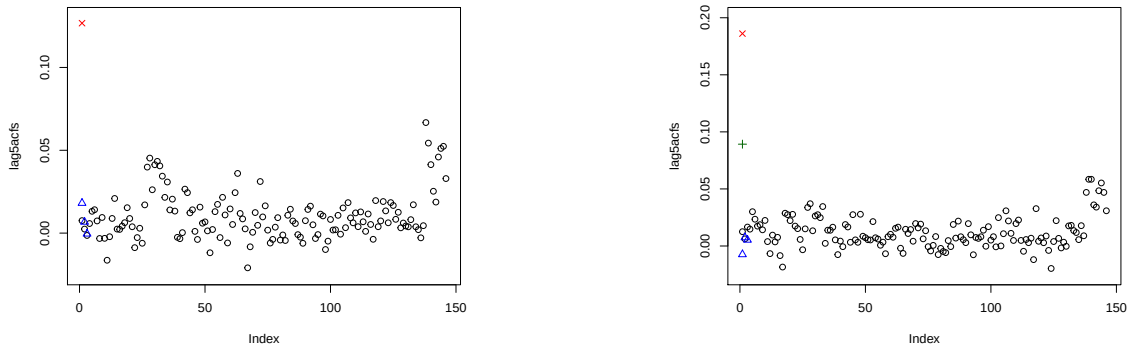
was and had a coefficient estimate of 3.75 (1.66, 5.80). We can see from Figure 7a that some LSOAs have higher random effects than others, so it may be worth fitting a model that also has spatially correlated random effects.

Our chosen spatial extra Poisson variation model has the form:

$$\log(R_i) = \log(pop_i) + \alpha + \beta_1 x_{\text{Employment},i} + \beta_2 x_{\text{Education},i} + U_i + S_i, \tag{7}$$

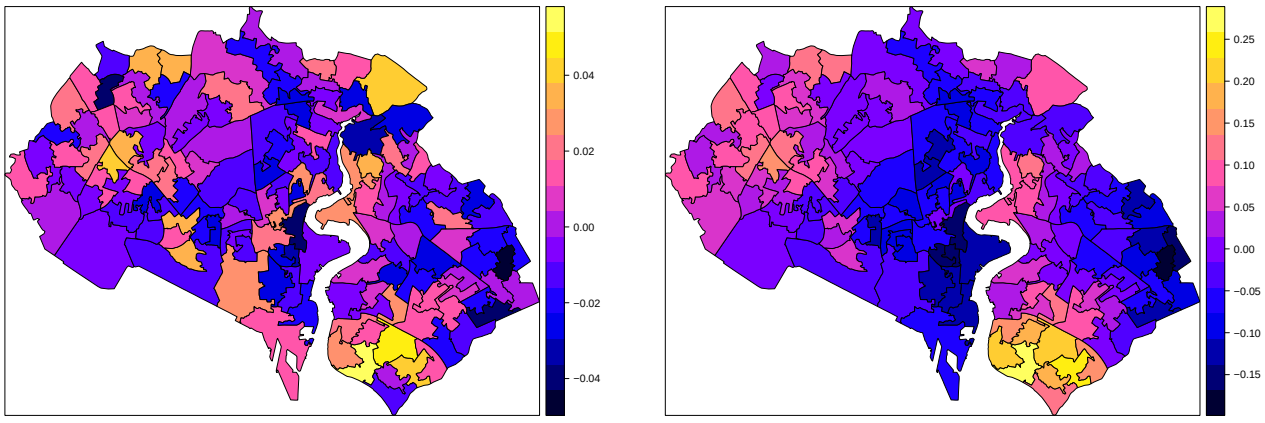
where α is the intercept, the β's are the coefficients of the covariates $x_{\bullet,i}$, U_i are mutually independent random effects without any spatial structure, and S_i are the spatially correlated random effects. All the parameter trace plots seem to be mixing well (omitted), and we can see from Figure 6b that, at lag 5, the ACF values for the parameters are all reasonable. The variances for the random components may again be a little high, but over all its reasonable. Again, under this model Education was marginally non-significant, with a coefficient of -0.005 (-0.0106, 0.0003), but Employment was significant 3.894 (1.6882, 6.1326). From Figure 7b, however, we can see there is still an area in the south-east which has much higher random effects than the rest of Southampton. Comparing the two plots in Figure 7 we can see that adding the spatially correlated random effects explains a lot of the non-spatial random effects variation, but there is still an area in the south-east that is much higher. Also, looking at the scales of the plots, we can see that the size of the random effects in the spatial model are much larger than in the non-spatial model. More formally, we can compare the models using the

Deviance Information Criterion (DIC), the model with the lower DIC is considered to be the best fitting, most parsimonious, model. The DIC for the non-spatial model is 713.57, and for the spatial model it is 703.77, so we say that, of the two models, the spatial model explains the data the best.



(A) (B)

FIGURE 6. (A) A plot of the lag 5 ACF values for the different parameters fitted in the independent extra Poisson variation model. The blue triangles are the fixed effects, β, the black circles are the random effects, ϕ, and the red cross is the variance parameter governing the variance of the random effects, τ^2. (B) A plot of the lag 5 ACF values for the different parameters fitted in the spatially-correlated extra Poisson variation model. The blue triangles are the fixed effects, β, the black circles are the random effects, ψ, which represent both the spatially and non-spatially random effects, the red cross is the variance parameter governing the variance of the non-spatial random effects, τ^2, and the green plus is the variance parameter governing the variance of the spatial random effects, σ^2.



(A) (B)

FIGURE 7. (A) A plot of the random effects for the model with only non-spatially correlated random effects. (B) A plot of the random effects (combined spatially correlated and independent random effects) for the model with both independent and spatially correlated random effects.

7. DISCUSSION

From all of our analyses it is clear that there exists unexplained spatial variance for the number of cases of non-specific gastrointestinal infections in each LSOA in Southampton. This

suggest that there are other variables that need to be adjusted for in our model, such as recent foreign travel [7]. In particular, Figure 7 shows that the random effects are quite large, especially once one includes spatially correlated random effects. The Index of Multiple Deprivation seems to explain a small amount of the variance in the model, but a subset of its components seems to be slightly better, but even then, it is clear that there are other, more important, factors that are missing from the model. Ideally, we would communicate our findings to scientists with knowledge of the subject, confirm that our variable selection is valid and supported scientifically, and inquire as to whether there are any other known variables that could explain the variation in risk.

Given the coefficients in our GAM and extra Poisson variation models, it would suggest that higher levels of deprivation, or rather, higher levels of certain characteristics that contribute to deprivation, lead to higher risk of non-specific GI infections. For instance, in our chosen extra Poisson variation model with spatially correlated random errors, Employment had a coefficient of 3.894 (1.6882, 6.1326), meaning that if an area has a high proportion of its working-age population involuntarily excluded from the labour market, the number expected number of cases of non-specific GI infections in that LSOA will increase. This is contradictory to some of the literature [7;10], but not all [10]. One possibility, however, is that these relationships would be removed once one adjusts for other risk factors, such as recent foreign travel [7].

In addition to including more informative variables, we could also improve our models by collecting individual level data, as well as data on real controls, to fit a binary model. We have used overlay operations to move to the aggregated data to the individual level, but true individual level data would be better, and we are also basing our analysis on simulated controls with the same aggregated data. Individual level data, however, can be very difficult to obtain, and comes with a lot more data protection issues, so it may be difficult to improve the models in this way. One improvement that would be possible, however, is to incorporate a temporal aspect into the model, as well as a spatial component, as was done in Diggle et. al (2005) [2]. The data already comes with the date of reports of cases, so this is a possible improvement.

In conclusion we have done some exploratory analysis to suggest some relationships that may exist between the level of deprivation in Southampton LSOAs and non-specific gastrointestinal infection, and how risk is distributed across the area. We then fit Poisson generalised additive models to the aggregated data to try and identify both LSOAs that had increased risk, and some important variables that played a role. In particular, we found that using a subset of the domains of the Index of Multiple Deprivation, Education and Employment, explained the variation better than the IMD score. There was still a lot of unexplained variance however. Due to this spatial variation, we decided to fit extra Poisson variation models, after discovering that the standard errors in a Poisson GLM were too small. We fit both non-spatial and spatial extra Poisson variation models using the variables determined for our GAMs, and using the DIC chose the spatial model to be better fitting. The spatial model, however, had large random effects, suggesting that there are other variables that we have not considered. Our models suggest that higher levels of deprivation are associated with higher counts of non-specific GI infection, though there is a lot of unexplained variation, and these relationships may be removed once one adjusts for the correct variables. Finally we suggest some improvements that could be made to our models, such as using individual level data or a temporal component.

References

[1] Sofie BieringSrensen, Grethe Sndergaard, Karen Vitting Andersen, AnneMarie Nybo Andersen, and Laust Hvas Mortensen. Time trends in socioeconomic factors and risk of hospitalisation with infectious diseases in preschool children 19852004: a danish registerbased study. *Paediatric and Perinatal Epidemiology*, 26(3):226–235, 2012.

[2] Peter Diggle, Barry Rowlingson, and Tingli Su. Point process methodology for online spatiotemporal disease surveillance. *Environmetrics*, 16(5):423–434.

[3] Peter J. Diggle, Leo Knorr-Held, Barry Rowlingson, Ting li Su, Peter Hawtin, and Trevor N. Bryant. *On-line monitoring of public health surveillance data.*, pages 233–266. Oxford University Press, 2004.

[4] Food Standards Agency. The second study of infectious intestinal disease in the community (IID2 Study). `https://www.food.gov.uk/science/research/foodborneillness/b14programme/b14projlist/b18021`, 2016. [Online; accessed 2018-03-21].

[5] H. Graham. *Understanding health inequalities*. Open University Press, 2009.

[6] Health Protection Surveillance Centre (HPSC). Infectious Intestinal Disease: Public Health and Clinical Guidance. `http://www.hpsc.ie/a-z/gastroenteric/gastroenteritisoriid/guidance/File,13492,en.pdf`, 2012. [Online; accessed 2018-03-20].

[7] G. J. HUGHES and R. GORTON. Inequalities in the incidence of infectious disease in the north east of england: a population-based study. *Epidemiology and Infection*, 143, 2015.

[8] Ministry of Housing, Communities and Local Government. English indices of deprivation 2015. `https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015`, 2015. [Online; accessed 2018-03-20].

[9] Nature. Gastrointestinal diseases. `https://www.nature.com/subjects/gastrointestinal-diseases`, 2018. [Online; accessed 2018-03-20].

[10] K. L. NEWMAN, J. S. LEON, P. A. REBOLLEDO, and E. SCALLAN. The impact of socioeconomic status on foodborne illness in high-income countries: a systematic review. *Epidemiology and Infection*, 143(12):24732485, 2015.

[11] NHS Inform Scotland. Gastroenteritis. `https://www.nhsinform.scot/illnesses-and-conditions/stomach-liver-and-gastrointestinal-tract/gastroenteritis`, 2018. [Online; accessed 2018-03-20].

[12] Tanith C. Rose, Natalie Adams, David C. Taylor-Robinson, Benjamin Barr, Jeremy Hawker, Sarah O'Brien, Mara Violato, and Margaret Whitehead. Relationship between socioeconomic status and gastrointestinal infections in developed countries: a systematic review protocol. *Systematic Reviews*, 5(1):13, Jan 2016.

[13] J C Semenza. Strategies to intervene on social determinants of infectious diseases. *Eurosurveillance*, 15(27), 2010.

[14] World Health Organisation. Social determinants of health: the solid facts. 2nd ed. `http://www.euro.who.int/__data/assets/pdf_file/0005/98438/e81384.pdf`, 2003. [Online; accessed 2018-03-20].

[15] World Health Organisation. Concepts and principles for tackling social inequities in health: levelling up part 1. `http://www.euro.who.int/__data/assets/pdf_file/0010/74737/E89383.pdf`, 2006. [Online; accessed 2018-03-20].

8. Appendix

```
### Density Plots ###

qplot(IMD, data = cvdata, geom = "density",
      fill = flag, alpha = I(0.5), main = "Distribution of cases by Deprivation",
      xlab = "IMD", ylab = "Density") + scale_fill_discrete(name="Subject")
# Very little difference, slightly more cases than controls at higher IMD
# Lower levels of deprivation assocaited with more cases but also more controls

### K-function estimation ###

k1 <- Kest(Cases)
r1 <- k1$r # The automatically chosen radii to compute the K-function at
kest1 <- k1$iso # The estimate of the Kest function at different radii for isotropic method

# Comparing clustering in cases and controls:
Contrls <- ppp(x = controls[,1], y = controls [,2], window = win)
```

```
17
18 kcontrls <- Kest(Contrls, r = r1)
19 D <- kest1 - kcontrls$iso
20
21 x <- c(Cases$x, Contrls$x)
22 y <- c(Cases$y, Contrls$y)
23 cc <- c(rep("case", Cases$n), rep("control", Contrls$n)) #stick 'em all together
24
25 for (i in 1:nsims) {
26   cc <- sample(cc)
27
28   simcases <- ppp(x = x[cc == "case"], y = y[cc == "case"],
29                   window = win)
30   simcontr <- ppp(x = x[cc == "control"], y = y[cc == "control"],
31                   window = win)
32   dsim <- Kest(simcases, r = r, correction = "isotropic")$iso - Kest(simcontr, r = r, correction = "
       isotropic")$iso
33   Dsim <- cbind(Dsim, dsim)
34 }
35
36 qts1 <- apply(Dsim, 1, quantile, probs = c(0.025, 0.975))
37
38 plot(NULL, xlab = "r", ylab = "Estimated D function", xlim = range(r),
39      ylim = range(Dsim))
40 polygon(c(r, rev(r)), c(qts1[1, ], rev(qts1[2, ])), col = "lightgrey",
41         border = NA)
42 lines(r, D)
43 abline(h = 0, col = "red", lty = "dashed")
44
45 ### Intensity Estimation ###
46
47 denCases <- density(Cases, positive = TRUE)
48 # Don't specify bandwidth, uses simple rule of thumb
49 denCases
50 attr(denCases, "sigma")
51
52 plot(denCases, main = "Estimate of the intensity of process across the window")
53
54 # Can calculate bandwidth using cross validation methods
55 den3a <- density(Cases, sigma = bw.diggle, positive = TRUE) # Diggle method
56 attr(den3a, "sigma") # x2 = Calculated bandwidth # 103.17
57 den3b <- density(Cases, sigma = bw.ppl, positive = TRUE) # People method
58 attr(den3b, "sigma") # 169.9483
59
60 plot(den3a, main = "Estimate of the intensity of process across the window")
61 plot(den3b, main = "Estimate of the intensity of process across the window")
62
63 ### Generalised Additive Model ###
64
65 library(mgcv)
66
67 IndData <- cbind(cvdata, rbind(cbind(x,y), controls))
68 colnames(IndData)[19:20] <- c("xcoord", "ycoord")
69
70 fit <- gam(flag ~ s(xcoord, ycoord), data = IndData, family = binomial(link = logit))
71
72 plot(win, main = "Contour plot of chance of being case as opposed to control")
73 vis.gam(fit, view = c("xcoord", "ycoord"), n.grid = 100, plot.type = "contour",
```

```
74          add = TRUE, type = "response", color = "terrain",
75          too.far = 0.08)
76 plot(win, add = TRUE)
77 axis(1)
78 axis(2)
79
80 ## Aggregated data, Poisson GAM ##
81
82 fitPoi4 <- gam(count ~ offset(log(pop)) + Employment
83                + Education + s(xcoord, ycoord), data = LSOAdata, family = poisson(link = "log"))
84 summary(fitPoi4)
85
86 plot(win, main = "Contour plot of estimated number of cases")
87 vis.gam(fitPoi4, view = c("xcoord", "ycoord"), n.grid = 100, plot.type = "contour",
88          add = TRUE, type = "response", color = "terrain",
89          too.far = 0.08)
90 plot(win, add = TRUE)
91 axis(1)
92 axis(2)
93
94 FitPoiAIC <- AIC(fitPoi, fitPoi2, fitPoi3, fitPoi4)
95
96 ### Bayesian Model for aggregated data ###
97
98 library(rgdal)
99 library(spdep)
100 library(CARBayes)
101
102 ### Non-spatial extra Poisson variation model ###
103
104 ireg <- S.CARleroux(count ~ Employment + Education + offset(log(pop)), family = "poisson",data =
      shamp@data, burnin = 100000, n.sample = 1000000, fix.rho = TRUE, rho = 0, W = W, thin = 50)
105
106 ## Plotting the random effects
107
108 shamp$U <- apply(ireg$samples$phi, 2, mean)
109 spplot(shamp, "U")
110
111 # The areas to the south east and north east boarders seem to have higher random effects compared with
      those elsewhere. It might be sensible to fit a model that uses spatially correlated random effects.
112
113 ### Spatial extra-Poisson variation model ###
114
115 # we are aiming to pass the neighbour information into our model that includes a spatially correlated
      random effect.
116
117 sreg <- S.CARbym(count ~ Employment + Education + offset(log(pop)), family = "poisson", data =
      shamp@data, W = W, burnin = 100000, n.sample = 1000000, thin = 50)
118
119 shamp$Uspat <- apply(sreg$samples$psi, 2, mean)
120 spplot(shamp, "Uspat")
121 # The areas to the south east and north east boarders seem to have far higher random effects compared with
      those elsewhere, and is much more concentrated that in the model with just Ui
122
123 ireg # DIC = 713.5744
124 sreg # DIC = 703.7657
125 # So the model with spatially correlated random effects is the preferred model.
```