Rodrigo Arriaza
Arvin Rastegar
Johanna Weiss

Project Lattice Data

Spatial Epidemiology
September 21, 2022

Question 1

A first representation of the Standardized Mobility Rate (SMR) in the districts of Mersey and West Lancashire in the North-East of England from 1982 to 1991 can be seen in 1. Overall, the SMR ranges from a minimum of 0 to a maximum of 5.34, with a mean of 0.99 and a slightly lower median of 0.87, which assumes that the data is right-skewed. The SMR seems to be higher in the South and West and lower in the northern part of the map.

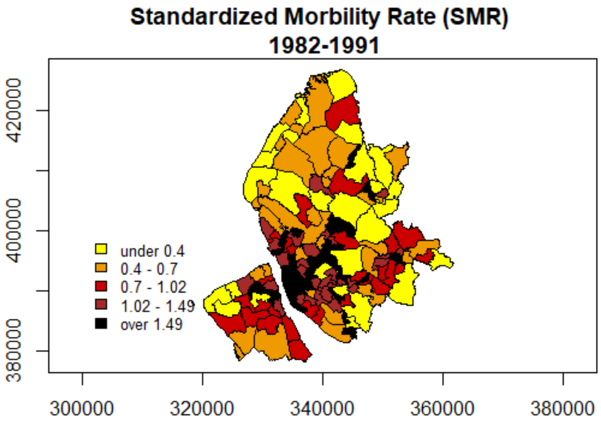


Figure 1: SMR in the districts of Mersey and West Lancashire

In Figure 2, the most and least connected regions are shown. In this analysis, neighboring regions are defined as the ones that share a geographical border. The most connected region (region number 88) has 16 borders with other regions and the 10 least connected regions all have two borders (regions number: 50, 64, 65, 76, 78, 94, 103, 110, 128, 144).

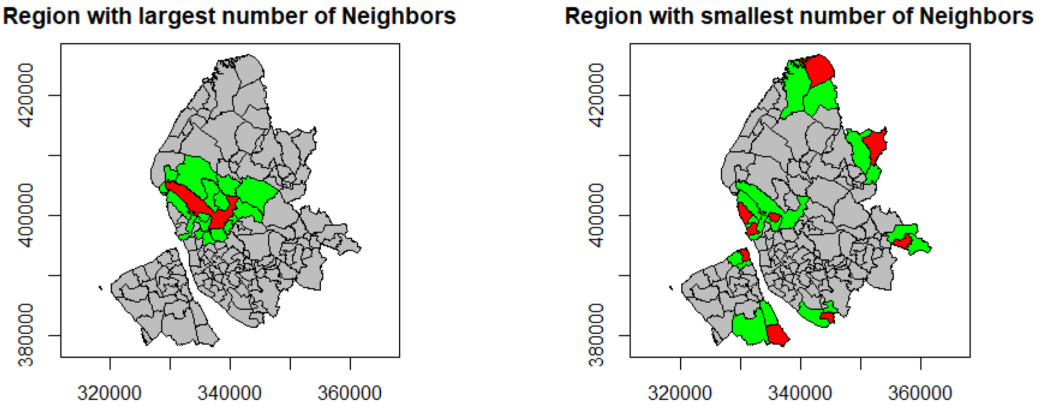

Region with largest number of Neighbors
420000
400000
380000
320000
340000
360000
Region with smallest number of Neighbors
420000
400000
380000
320000
340000
360000

Figure 2: Most and least connected regions

Spatial Correlation

In order to check the data for spatial correlation, Moran's I and the Geary's C indicators are calculated. The weights matrix that is used for the tests are based on the contiguity for each region and standardized with respect to the rows of the matrix. Moran's test is investigating global spatial correlation using the following hypothesis:

H_0 : No spatial correlation
H_a : Positive spatial correlation

The assumption for performing the test is that the data is approximately normally distributed. In order to meet this criterion the data is transformed using the Freeman-Tukey transformation. Performing the Shapiro-Wilk normality test shows that there is no evidence against normality of the data, therefore, Moran's test can be used.
Moran's I test under randomisation gives a p-value of < 0.001 and an I statistic of 0.29. As the p-value is lower than the significance level of 0.05 and the I statistic is positive, it can be concluded that there is global positive spatial correlation, as H_0 is rejected. Additionally, Moran's test using Monte Carlo permutations was performed, resulting in the same statistics and conclusion.
To determine if the data is also locally spatial correlated, Geary's test is performed, using the same hypothesis as above. This results in a C statistic of 0.69 and a p-value of < 0.001. Again, the H_0 hypothesis is rejected, which indicates positive local spatial correlation. This result is as well obtained using Geary's test with Monte Carlo simulations.

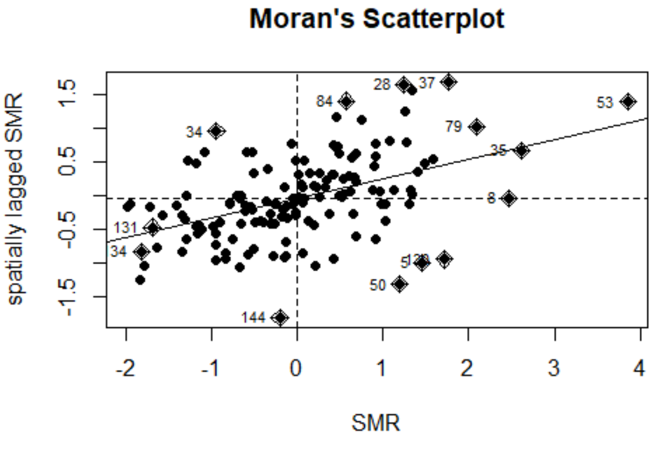


Figure 3: Moran's scatterplot

Figure 3 shows the Moran scatter plot of the data. It shows the standardized SMR on the x-axis and the spatially lagged values on the y-axis, so the spatially weighted sum of a region's neighbors. As there is a positive trend in the plotted data, the results obtained by the test, namely that the SMR data is positively spatial correlated. The regions that have the highest correlations with respect to their neighbours are highlighted in the plot, such as region number 53, 37,36, and 134. Against the overall trend, there are also regions that have very different SMR compared to their neighbors and therefore not positively correlated, such as region number 34 or 50.

Clusters and Hotspots

In order to detect clusters and hotspots in the data, the local Moran test is used, which calculates an I statistic for each region based on the spatial weights. Figure 4 shows these local I indicators, grouped in 4 intervals. A negative local I statistic means that the data shows different values than their neighbors, so called hotspots, whereas high values for I show that regions have similar values as the neighboring ones, called clusters.

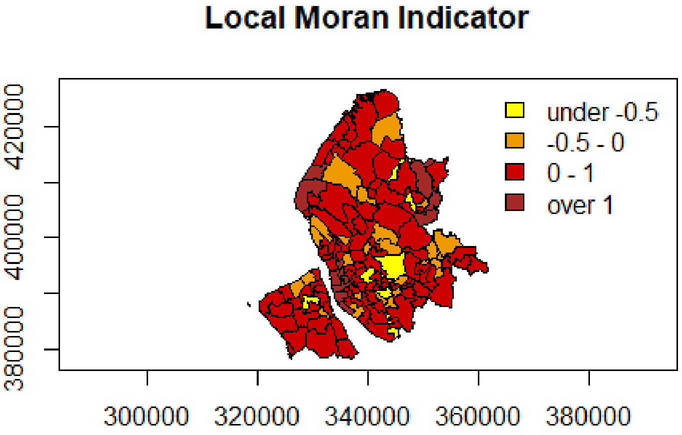


Figure 4: Local Moran Indicator

The figure shows for example a cluster in mid-west of the map, around Liverpool, where the regions are stronger correlated with their neighbours.

Question 2

Considering that the number of cases of Larynx follows a Poisson distribution, it is checked if the data is overdispersed. Overdispersion occurs when the observed variance in the data is higher than the variance of a theoretical model, in this case the poisson model. In order to check for overdispersion in the fitted model, the deviance and the degrees of freedom are looked at. The residual deviance is divided by the degrees of freedom, which gives the dispersion parameter, in this case, is 2.77. This can be interpreted as overdispersion in the data as is contains higher deviance than one would expect from a poisson model.

In order to check if different models give better results with repsect to its dispersion parameter, the data is fitted to a negative and quasi-poisson model.

Model	Negative Binomial	Poisson	Quasi-Poisson
AIC	752.92	862.15	NA
Dispersion	3.52	2.77	3.10

Table 1: Comparison of fitted models

The dispersion parameters for a quasi-poisson and negative binomial model can be found in Table 1. Additionally, the negative binomial and the poisson model are compared with respect to their AIC. The metric cannot be used to evaluate the quasi-poisson model, as quasi-poisson does not have a valid definition of a likelihood. The AIC parameter tells the best fit with the lower number, hence, the negative binomial model fits the data better.

Figure 5 shows the plots of three different models that have been fitted by the data. We use the residual plots to detect non-linearity, unequal error variances, and outliers. The residuals appear on the y-axis, and the predicted values appear on the x-axis. Ideally, residual values would be equal and randomly spaced around the horizontal axis. In all of the the residual plots, heteroscedastic data (data has different variances in different subsets of dataset) is detected. When data moves along the x-axis the distances along the horizontal axis increase. They also show non-linear patterns and outliers in the data sets. The other plots that are compared here are Q-Q plots. These plots help us assess if the residuals obtained after fitting the model plausibly follow a normal distribution. All three plots have more points above the line than normally distributed residuals would have, hence we could comfortably interpret that they are right-skewed. Overall, the negative binomial model performs best.

To conclude, the negative binomial model gives the best fit for this data. It might be due to overdispersion and the fact that negative binomial models are two-parameter poisson regression with an extra parameter for dispersion of the data.

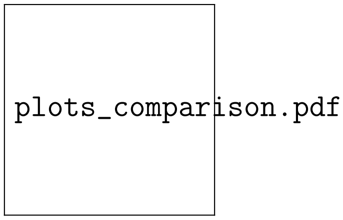


Figure 5: Our three models compared to each other

Question 3

We fit the SMR data considering the over-dispersion of the data using a heterogeneity model, a spatial (CAR intrinsic), and the convolution model. Then we estimate the models using Bayesian inference (Gibbs Sampling) using three chains of initial values.

In order to fit the SMR using these 3 models, we make use of the WinBUGS software to run a Bayesian Inference using Gibbs Sampling. We define the convolutional, spatial and heterogeneity models, load the data and the initial values for the priors and then we run the sampling until we see that the chains have converged.

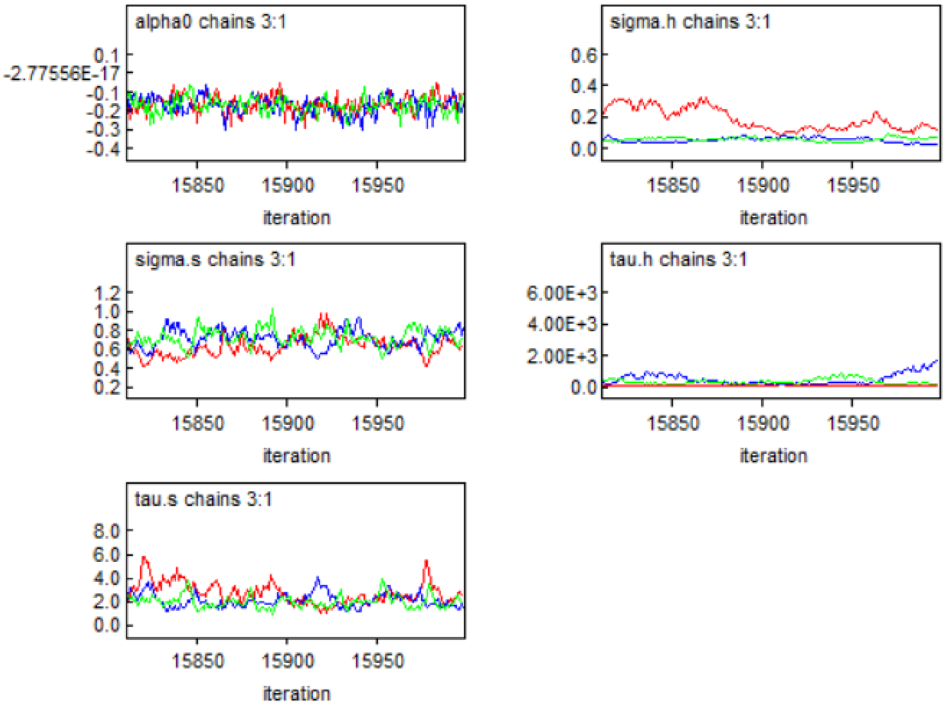


Figure 6: Convolutional model chains tracing

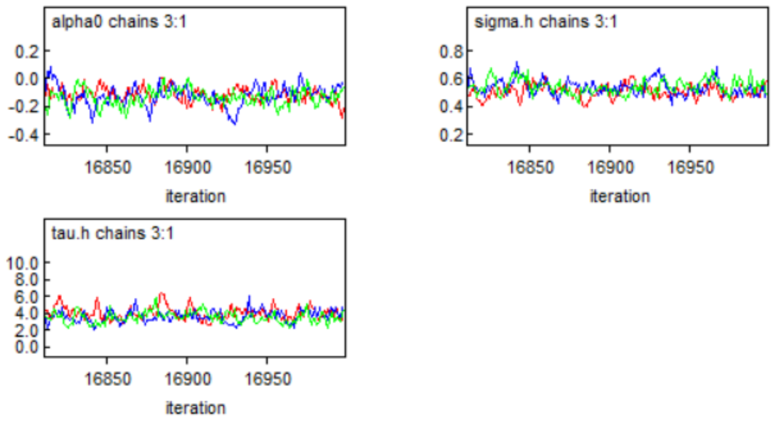


Figure 7: Heterogeneity model chains tracing

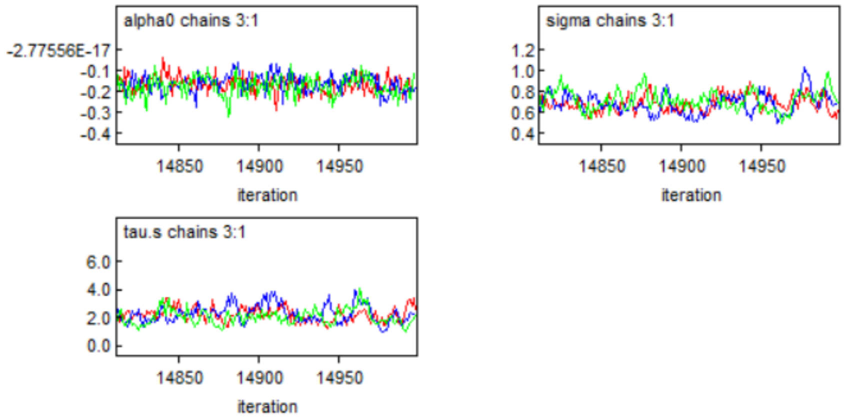


Figure 8: Spatial model chains tracing

Question 4

By checking the convergence of the chains, we can decide which model fits the data better.

As seen in Figures 6, 7 and 8, for every model the 3 chains are intertwined, which means they have converged. By checking the convergence of the chains, we can decide which model fits the data better.

To compare the models using the table, we use the DIC(Deviance Information Criterion) metric. It is the sum of 2 other parameters Dbar and pD. Dbar is the posterior mean of the deviance, pD is 'the effective number of parameters'. The model with the smallest DIC is estimated to be the model that would best predict a replicate dataset with the same structure as that currently observed.

Model	Dbar	pD	DIC
Spatial	618.75	56.04	674.09
Convolutional	616.15	58.86	675.05
Heterogeneity	616.16	80.69	697.21

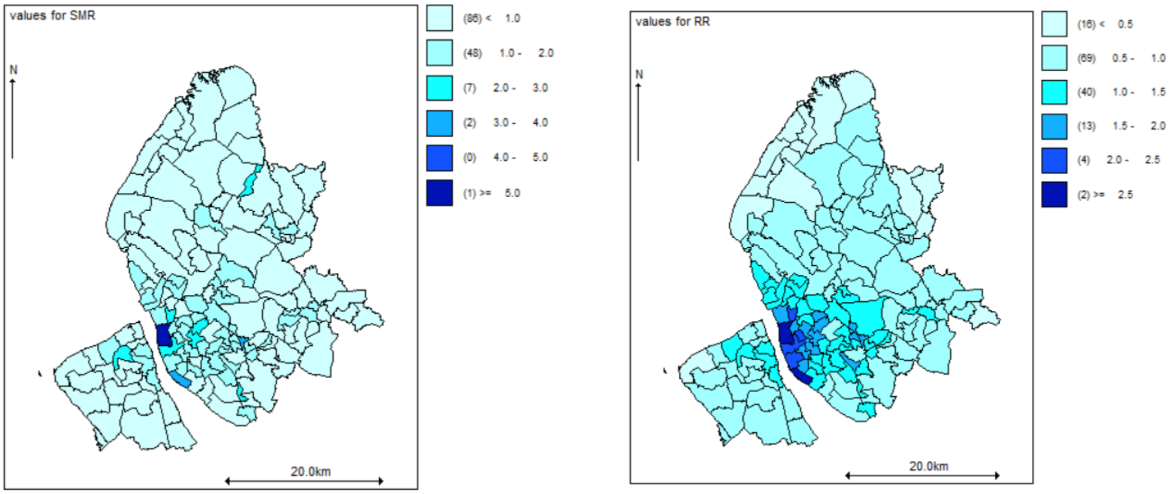
Table 2: DIC model comparison

As seen in Table 2, the model that best fits the data is the spatial model.

Question 5

As shown in Figure 9, in the SMR map, there are just a couple of regions with high values, which means that those are the ones where the observed cases surpass the expected ones, and therefore have a higher risk in comparison to the rest of the population.

On the other hand, for the relative risk plot, it is seen that the regions neighbouring the ones with higher SMR, will have a higher relative risk, meaning that the people there are more likely to suffer larynx cancer, since they are close to the more dangerous regions. However, for both values, the regions that are closer to the water have higher values for both SMR and RR, which could indicate that being in the proximity of the Mersey river increases the risk of suffering the disease.



(a) SMR mapping

(b) Relative Risk mapping

Figure 9: SMR and RR mappings using the spatial model

Figure 10 shows the posterior probability distribution of the relative risk. It can be seen that most values are either very close to 0, mainly in the northern part of the map; or close to 1, in the southern part, around the estuary of the river. Hence, the posterior distribution, estimates the risk higher in those regions than previously indicated by the model. Similarly, the areas that were considered low-risk regions before, are adjusted to even lower probabilities in the posterior probability distribution.

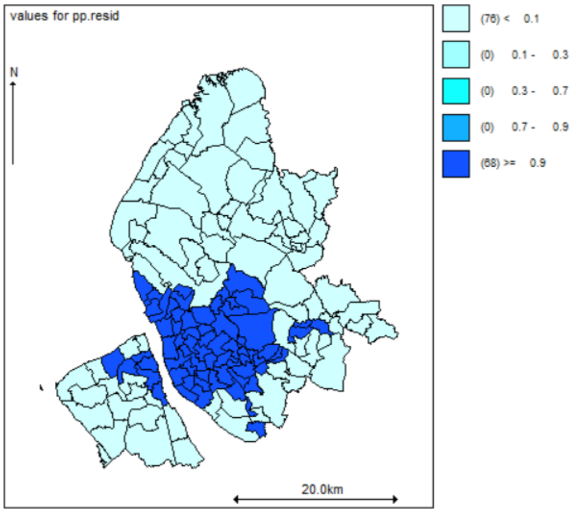


Figure 10: Posterior probability of the relative risk

Figure 11 shows the random effects of the spatial model. It can be seen that the regions that showed higher values for SMR according to the previous plots, are also the ones that assume more random effects contributing to the model.

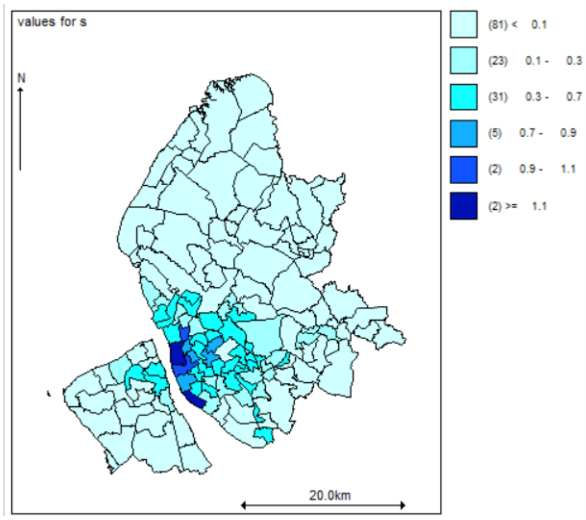


Figure 11: Random effects of the spatial model