Simulating the spread of Covid – 19, the UK case

Bernard Fingleton

Department of Land Economy, Silver Street, University of Cambridge, CB3 9EP, UK

Version 1

Very initial draft discussion paper. Comments and suggestions welcome

05/04/2020

Methodology

The starting point for this simulation is the distribution of the virus at as 31 March 2020. In fact data on the spatial distribution across English districts is published on the UK Government website https://www.gov.uk/government/publications/covid-19-track-coronavirus-cases. There is a breakdown by 149 Upper Tier Local Authorities (UTLAs) but the accuracy of these data is highly suspect. Additionally, the geography of the breakdown only partially coincides with the geography and mapping tools I have to hand, the problem being to convert the numbers in the UTLA breakdown to my own Unitary Authority and Local Authority Districts (UALADs, old definition) numbering 408 and covering the whole of Great Britain. Time is of the essence, and although conversion from one geography to another and filling in the missing data for Scotland and Wales is certainly doable, but it would be a tedious, boring and time consuming exercise based on dubious numbers. Rather, the approach adopted is to regress the Government numbers for some districts on their population size and population density, and this use the estimated coefficients to obtain predicted numbers of cases for districts for which I don't have data. The outcome is then scaled to equal the published UK Government total confirmed cases at the start date, summing across all areas this is 25,100 as at 31 March 2020.

Starting from these initial cases as at 31 March, which I call week 1, week by week numbers of cases across all 408 UALADs of GB are obtained using the following formulation, which is based on Baltagi et al(2014,2019).

$$\ln c_t = k + \gamma \ln c_{t-1} + \rho_1 \mathbf{W}_N \ln c_t + \beta_1 \ln p_t + \beta_2 \ln d_t + \theta \mathbf{W}_N \ln c_{t-1} + \varepsilon_t \quad (1)$$

In equation (1), c_t is the number of cases across $N =$ 408 districts at time t , $t = 1,...,10$ weeks. k is a constant, p is the population total in each UALAD d is the population density of each UALAD, which are assumed to be constant over the 10 week period of simulation. Note the suffixes allow lagged effects, so that the number of cases in a district depends on the number in the previous week, but it also depends on contemporaneous effects coming from nearby districts, as governed by $\mathbf{W}_N$, reflecting infectivity across space, with proximate districts having related numbers of cases. The number of cases in a district at time *t* also depends on the number in nearby districts in the previous week, reflecting infectivity across time and space. Of major importance is each district's population size and density. A high population total will naturally lead to a higher number infected, but density of population is also assumed to be equally important. People living in close proximity in a dense location will be less able to socially isolate and the probability of coming into contact with an infected person will be higher.

$\mathbf{W}_N$ is a (standardised) N by N connectivity matrix with zeros on the main diagonal. Districts are considered to be connected if they share a district boundary, that is they are contiguous. This gives an N by N matrix of 1s and 0s where 1 denotes contiguity and 0 otherwise. This is then row standardised by dividing each cell of the matrix by its row total, giving $\mathbf{W}_N$ in which rows sum to 1. So $\mathbf{W}_N \ln c_t$ is an N by 1 vector with cell i $(i = 1,...,N)$ equal to the weighted average, with equal weights, of c_t in districts that are contiguous to district i.

Other (unobservable) factors are captured by ε_t which is a vector of random effects picking up time-invariant heterogeneity across districts (the net effect of many social and economic factors), denoted by $\mu_i, i = 1,...,N$. In addition, it picks up idiosyncratic time and country varying shocks, denoted by ν_{it} .

Also as is utilised in recent literature (Baltagi et al, 2019), we assume a spatial moving average process causing error dependence across districts, thus

$$\begin{aligned} \varepsilon_t &= u_t - \rho_2 \mathbf{W}_N u_t \\ u_{it} &= \mu_i + \nu_{it}, i = 1,...,N, t = 1,...,T \\ \mu_i &\sim iid(0, \sigma_\mu^2) \\ \nu_i &\sim iid(0, \sigma_\nu^2) \end{aligned} \qquad (2)$$

The aim is to use the model to simulate realistic and plausible outcomes across districts and over time. Normally one would estimate the parameters $\gamma, \rho_1, \beta, \theta, \rho_2, \sigma_\mu^2$ and σ_ν^2 but in this case we simply assume values that hopefully make sense and correspond to what seems evident from data and other modelling exercises relating to coronavirus. Accordingly, we assume $k = 2, \gamma = 0.5, \rho_1 = 0.2, \rho_2 = -0.1, \beta_1 = 0.15, \beta_2 = 0.15, \theta = -0.25$. The behavioural assumption behind these parameters are very simple, they are chosen to ensure that the number infected is a function of spatial and temporal proximity. It will be a large number if the number in the district is large in the previous week, or if a large number of cases occurs in a nearby district, either in the same or in the previous week. Likewise, it will be large in large, dense centres of population.

In the interests of a happy outcome, these parameter assumptions are commensurate with a dynamically stable, stationary process. This means that in the long run the number of cases in each district tends to its own equilibrium level consonant with its population, population density and proximity to other districts. The literature cited gives the precise rules governing whether or not the dynamics reduce to a steady state over time. Of course the assumed parameter values are open to question, and different analysts might want to give more or less weight to different elements. For example, should ρ_1 be larger to reflect a stronger impact of the number of cases in neighbouring districts? The negative value of θ may seem to be at variance with the assumed positive proximity effect, but what we are trying to obtain is an overall net positive effect of proximity which, importantly, leads to an equilibrium outcome rather than runaway transmission. Baltagi et al(2019) and Fingleton and Szumilo(2019) give the logical basis for a negative θ parameter consistent with positive spatial dependence. I have avoided choosing parameters leading to exponential growth in the number of cases with no long run equilibrium. The model is therefore trying to produce an outcome that is in line with what the UK Government wants, and expects if UK citizens adhere strongly to the policy of social distancing.

The equilibrium outcome at time T, where T is a large number and for this rapidly evolving process is as short as 10 weeks, is given by

$$\ln c_T = \left(\mathbf{B}_N - \mathbf{C}_N\right)^{-1}\left(\beta_1 p + \beta_2 d + \mathbf{G}_N \mu\right) \quad (3)$$

In this, $\mathbf{B}_N = \left(\mathbf{I}_N - \rho_1 \mathbf{W}_N\right)$, $\mathbf{C}_N = \left(\gamma \mathbf{I}_N + \theta \mathbf{W}_N\right)$ and $\mathbf{G}_N = \left(\mathbf{I}_N - \rho_2 \mathbf{W}_N\right)$, p is the population and d is the population density at time T which is assumed to remain constant for each district over time, and μ is an N by 1 vector of time-invariant heterogeneity across districts. To generate μ we make the heroic assumption that they occur at random across districts by drawing at random from an N(0, 0.01) distribution. In practice, the evolution towards equilibrium is given by the recursive iteration through to time T of

$$\ln c_t = \mathbf{B}_N^{-1}\left(k + \mathbf{C}_N \ln c_{t-1} + \beta_1 p + \beta_2 d + \mathbf{G}_N \mu\right), t = 2,...,T \quad (4)$$

Where the process commences with $\ln c_{i1}, i = 1,...408$ is the assumed distribution of cases across 408 GB districts at 31 March 2020 as described above. This gives exactly the outcome at time T as equation (3). The reason for the iteration is that we wish to depict the dynamic evolution of the number of cases and the share of population infected through time until steady state has been reached.

Outcomes

Figure 1 shows that the number of cases peaks at around week 6, but as Figure 2 shows we should see a steady decline in the number of new cases each week as the lockdown takes effect. Figure 3 is the geography of cases in week 1 (31/3/2020 ,of course the numbers are subject to error!). The broad pattern is that places with large populations and high population densities have the highest incidence, of course reflecting the fact that the starting point for the simulation is in many districts based on their population and population density. The upper limit of about 3000 cases occurs in the highly urbanized districts (London, Midlands, Central belt of Scotland etc). Roughly half of districts have 1000 cases or less. Figure 5 focuses on the London Boroughs, which have seen the greatest intensity. Most Boroughs end up with between 10,000 and 15,000 cases by week 6. Figure 6 gives the geography of the log number of cases in week 1, with an intense concentration in Inner London, plus some more peripheral urban hotspots in the Greater South East. Figure 8 gives the official published cases by region in week 2 (as at 5/4/2020) which can be compared with the simulated number from week 2. The geography is similar, but the aggregate number adding across all 408 districts of Figure 9 is 1,498,200 which is very much larger than the official count summing over the Figure 8 regions, which is 47,800. Note however the caveat issued with the official total, which says that 'the actual number of people with the respiratory infection in the UK is estimated to be much higher though - as only those in hospital and some NHS staff are currently tested'. At the time of writing I have no reason to believe the number is not about 1.5 million. Figure 11 shifts attention to week 5, predicted to be towards the end of the epidemic (we hope!). The focus on the urbanized districts remains the same, but Figure 12 shows that the number of cases has greatly increased (paralleling the shift from week 1 to week 5 in Figure 1). Figure 13 quantifies the increase by district, showing that urban districts are expected to see 5 or 6 times the week 1 cases, compared with more peripheral and rural areas where the ratio is more like 4. Figure 15 gives the same for London and the South East, showing a distinction between urban and rural districts.

In Figure 17 attention switches to the share of the population infected. In week 1 the assumption is that few districts have above 2% of the population infected. By week 5 most remain at about 3% but some districts approach or exceed 8-10%. Figure 18 gives the geography of log infection rate in week 1. Evidently the hotspots for infection are not just highly urbanized areas, but a bit more widespread. Figure 19 indicates that about half of all districts have infection shares of 1% or less. Figures 20 and 21 gives a similar, persistent, geography but higher infection shares for week 5. Figure 22 shows the geography of the change in infection shares from weeks 1 to 5. The emphasis here is that the highest % change is in and around urbanized areas. Figure 23 indicates this is up to 6-8%, compared with roughly 3% for more rural districts. Figure 24 compared infection shares in London and Edinburgh. In London, they reach 5%, but it is less than half that for Edinburgh.

Figure 1 : Cases week 1

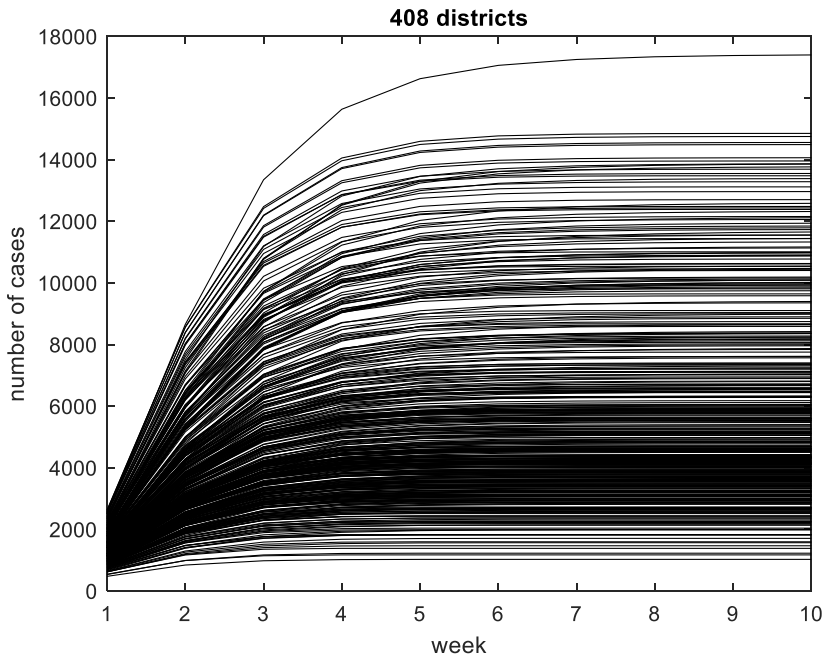


Figure 2 : Change in number of cases

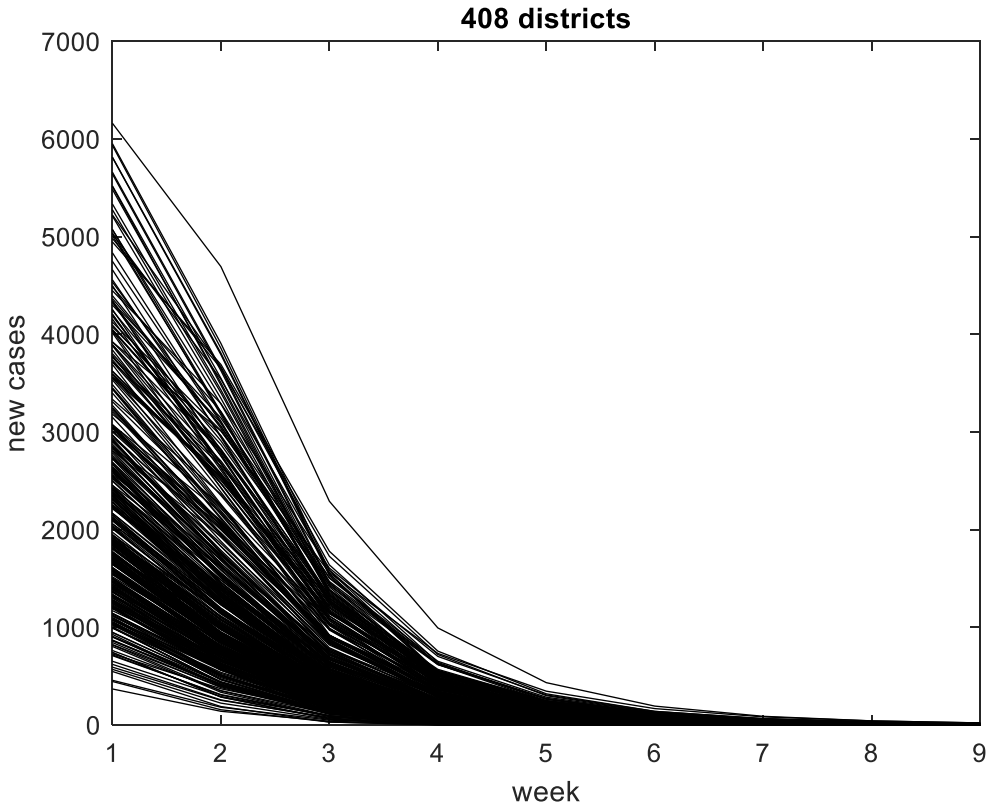


Figure 3 : log cases week 1

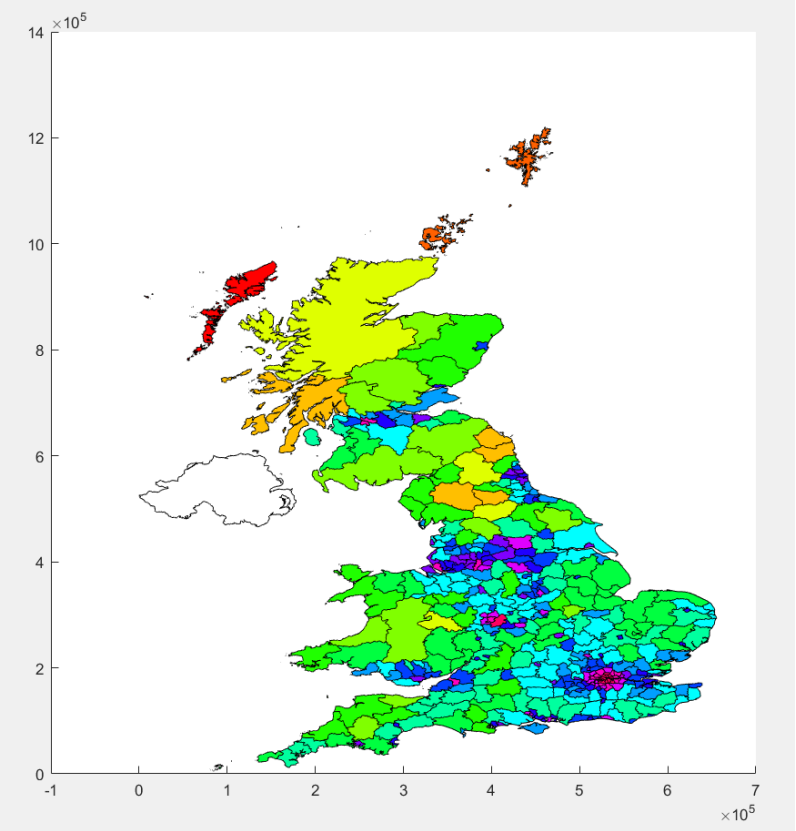


Figure 4 : Frequency distribution log cases week 1

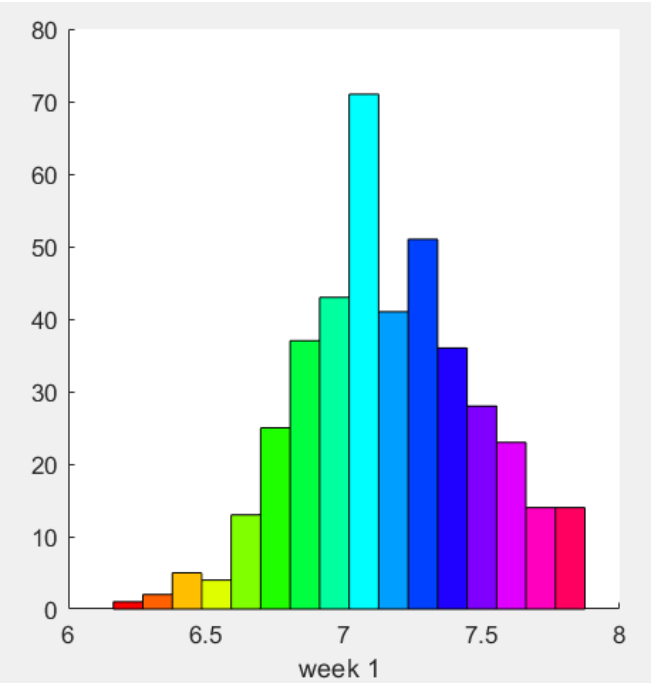


Figure 5 : London Boroughs cases week 1

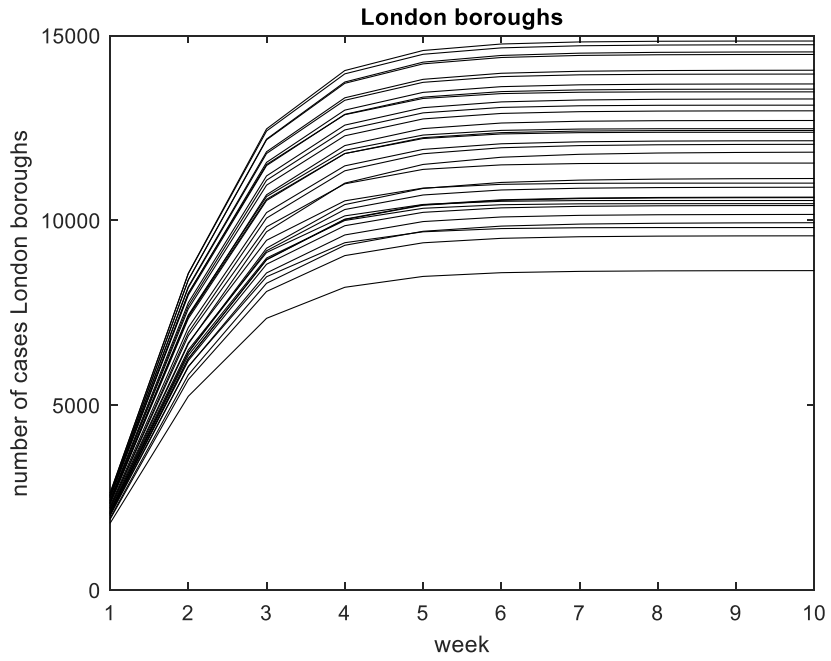


Figure 6 : London and the South East log cases week 1

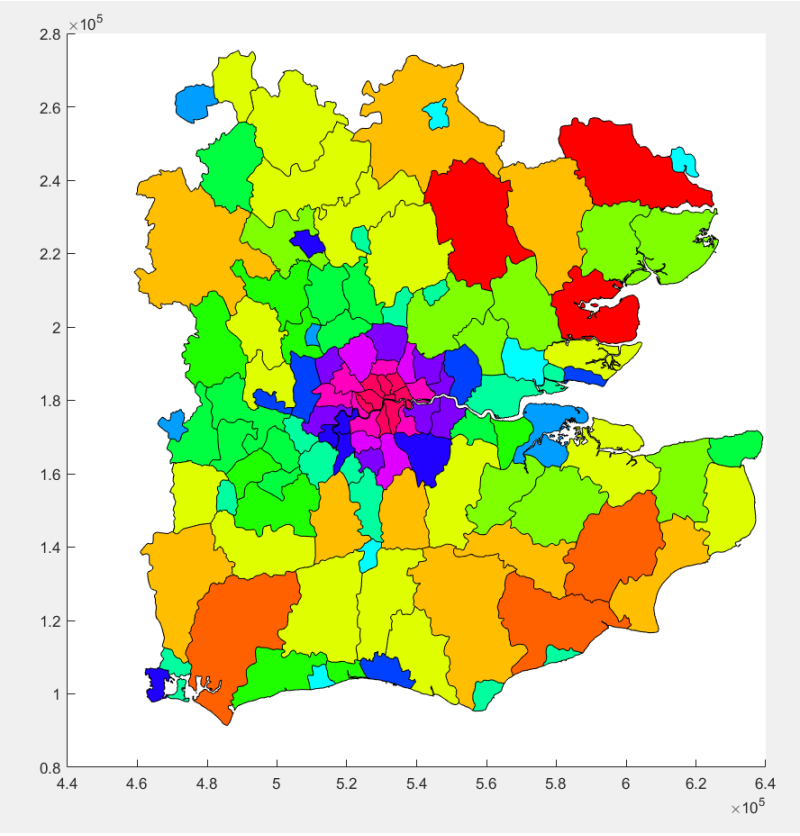


Figure 7 : Frequency distribution log cases London and the South East week 1

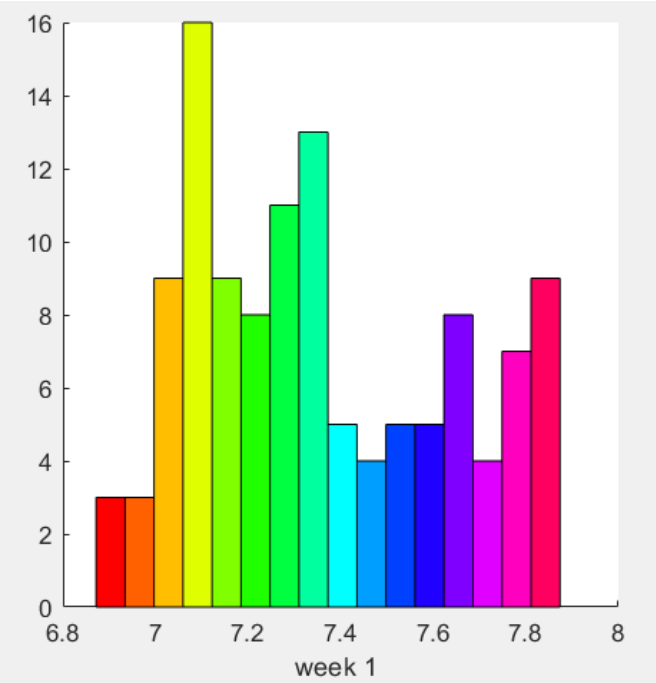


Figure 8 : Confirmed cases week 2

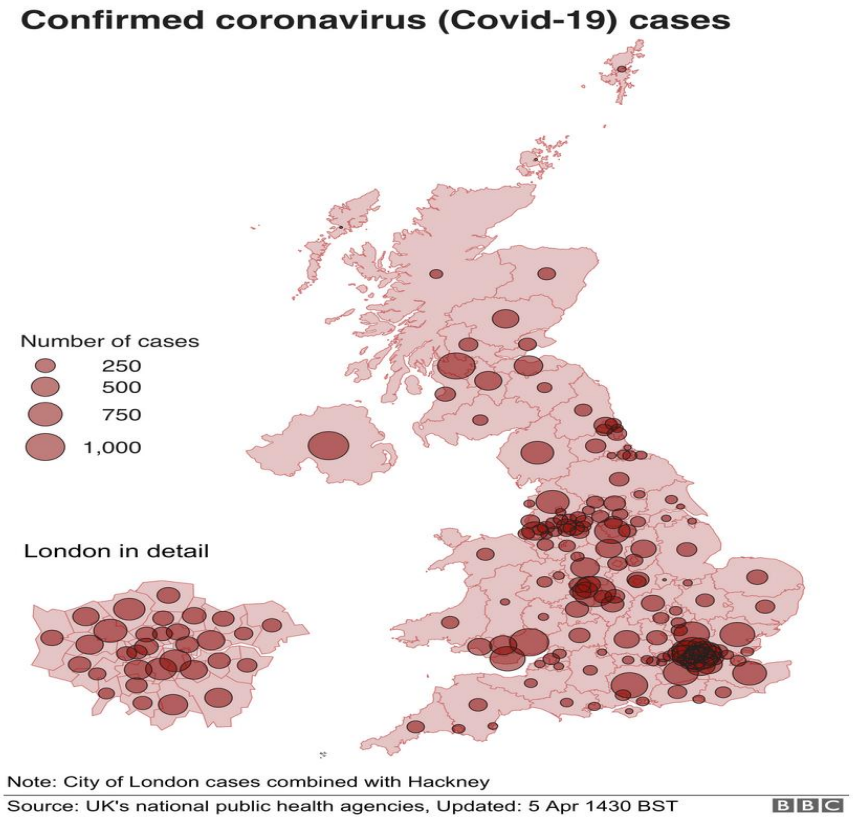


Figure 9 : simulated cases week 2

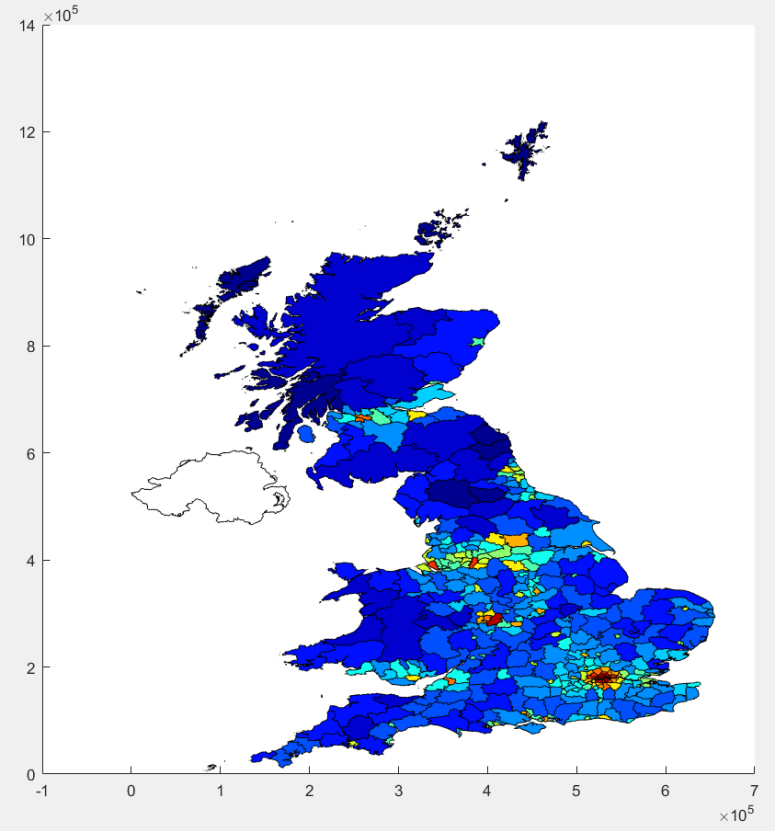


Figure 10 : Frequency distribution of number of cases week 2

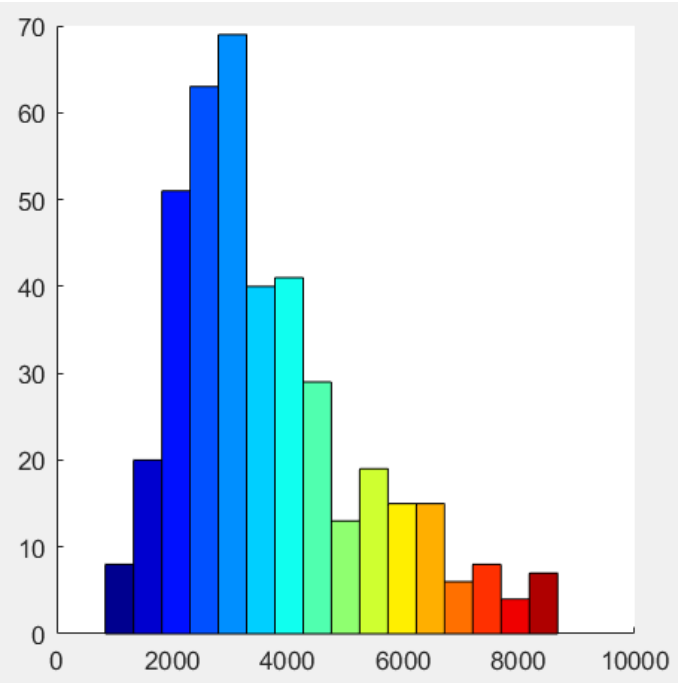


Figure 11 : log cases week 5

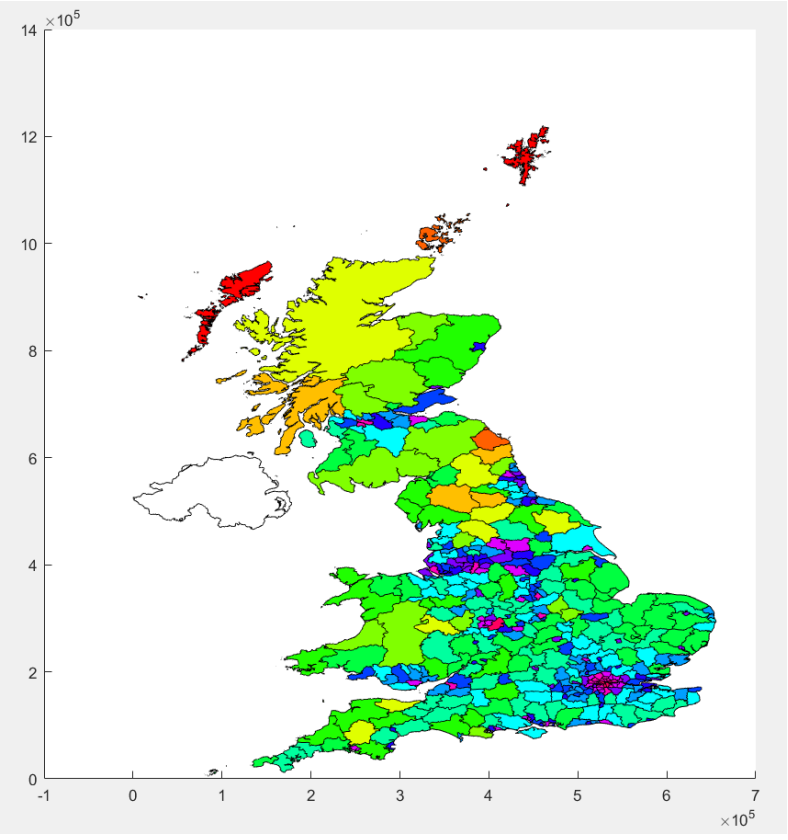


Figure 12 : Frequency distribution log cases week 5

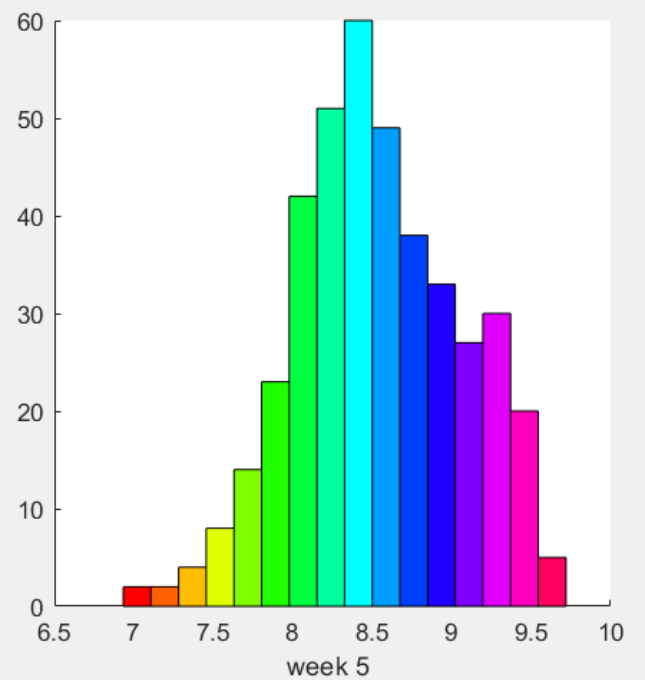


Figure 13 : ratio of cases week 5 to week 1

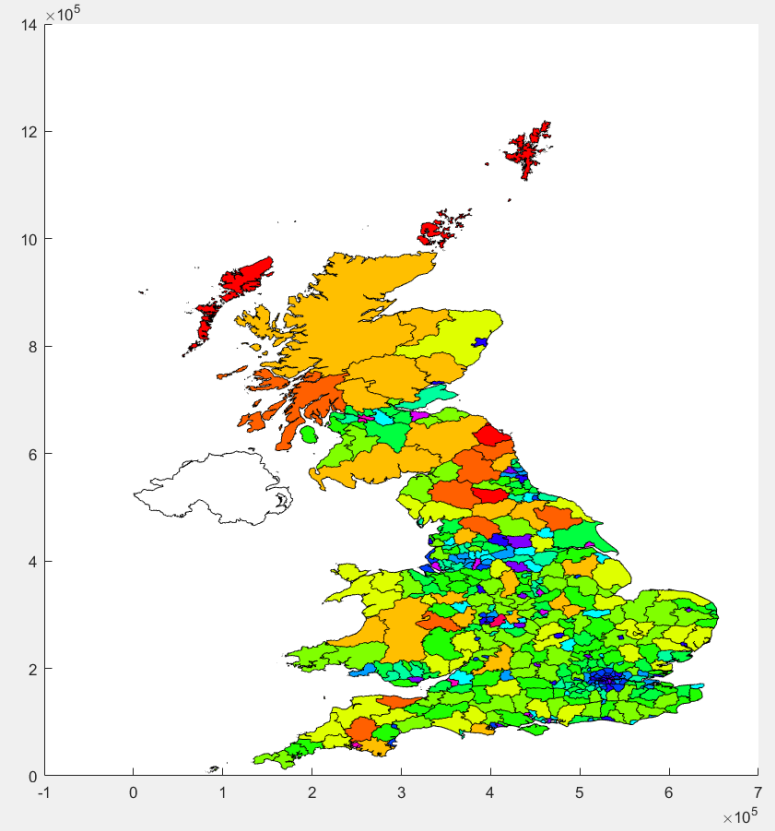


Figure 14 : frequency distribution of ratios week 5 to week 1

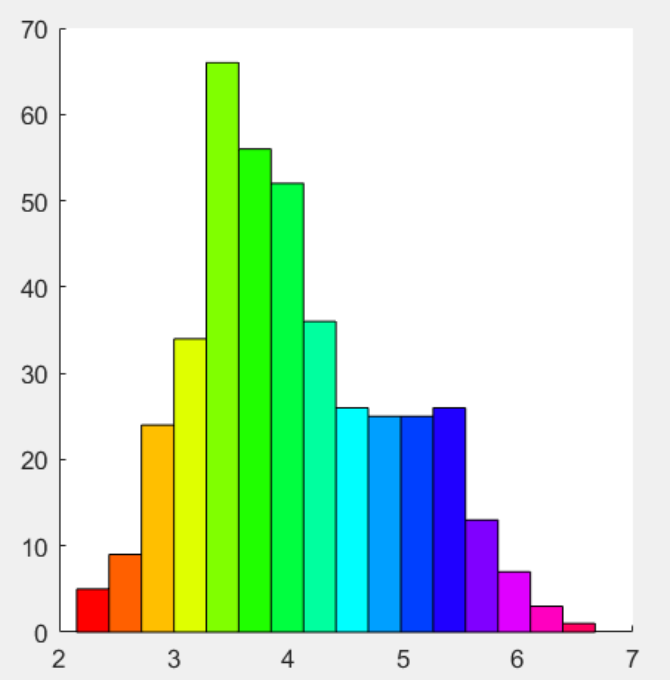


Figure 15 : ratio of cases London and the South East week 5 to week 1

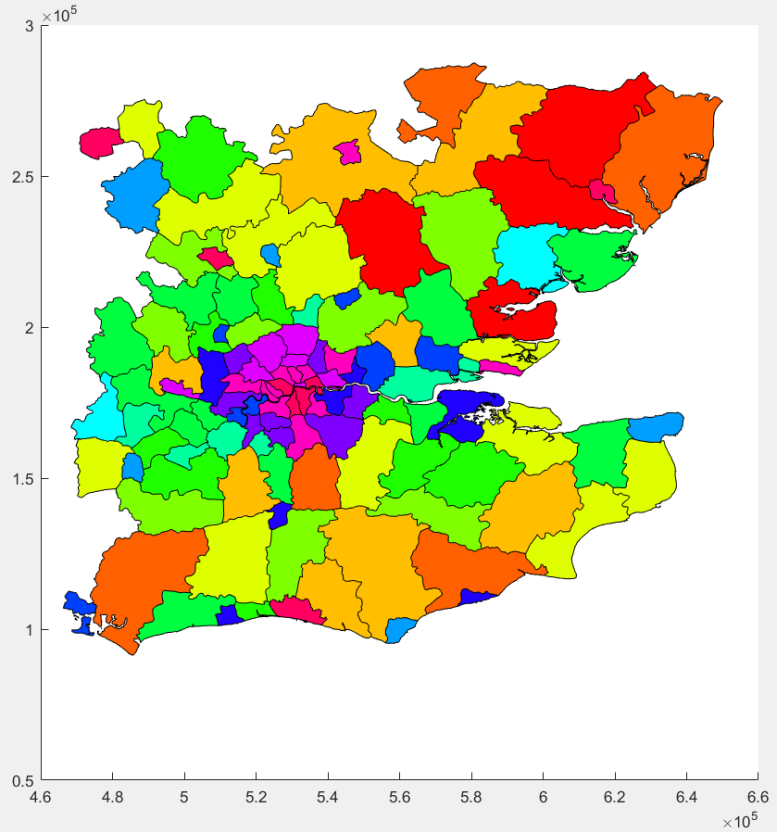


Figure 16 : Frequency distribution ratio of cases London and the South East week 5 to week 1

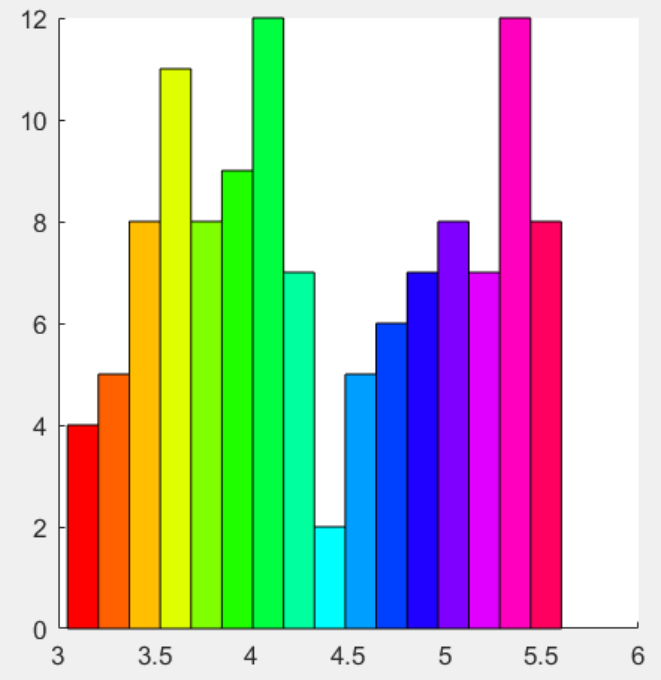


Figure 17 : Infection share by week

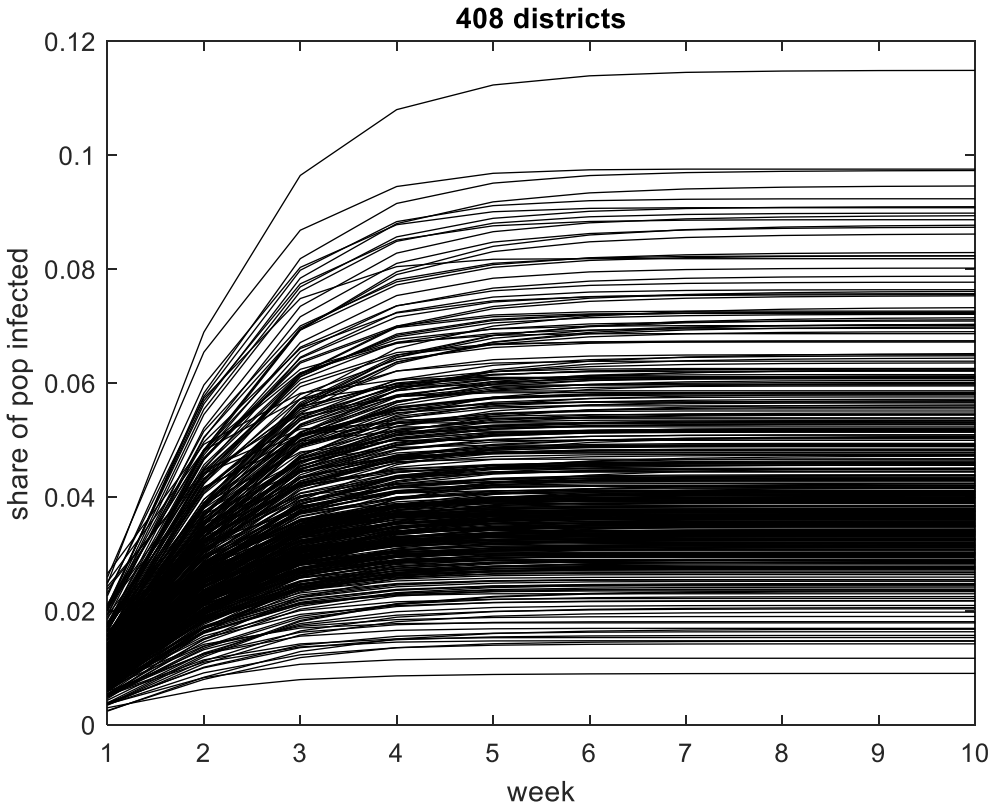


Figure 18 : Log infection share week 1

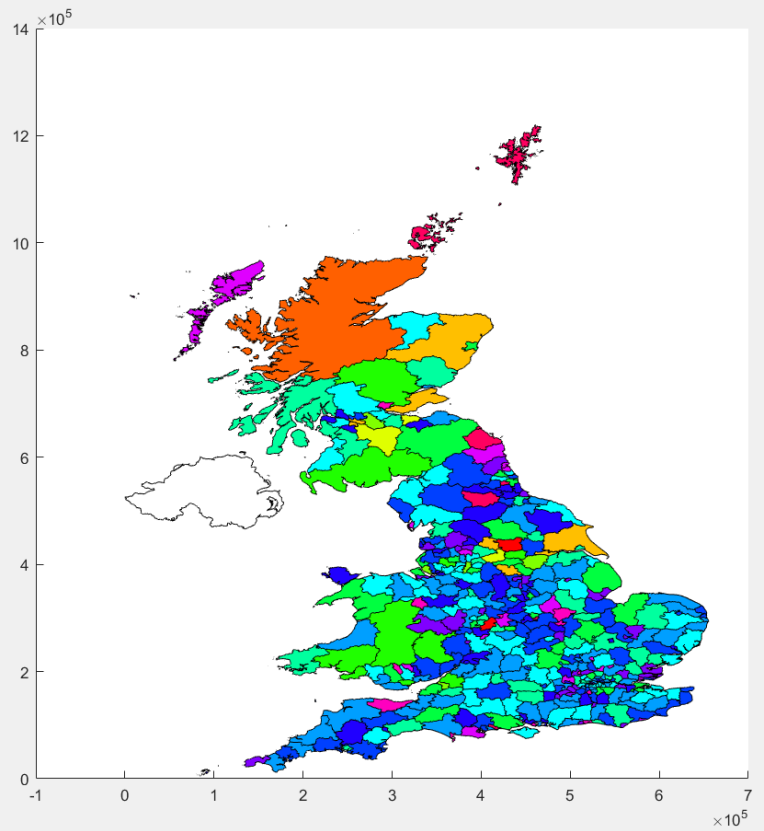


Figure 19 : Frequency distribution Log infection share week 1

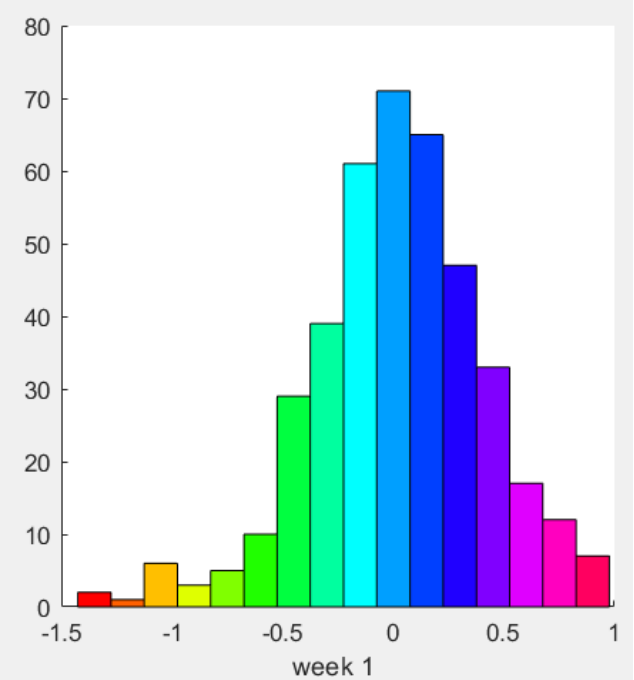


Figure 20 : Log infection share week 5

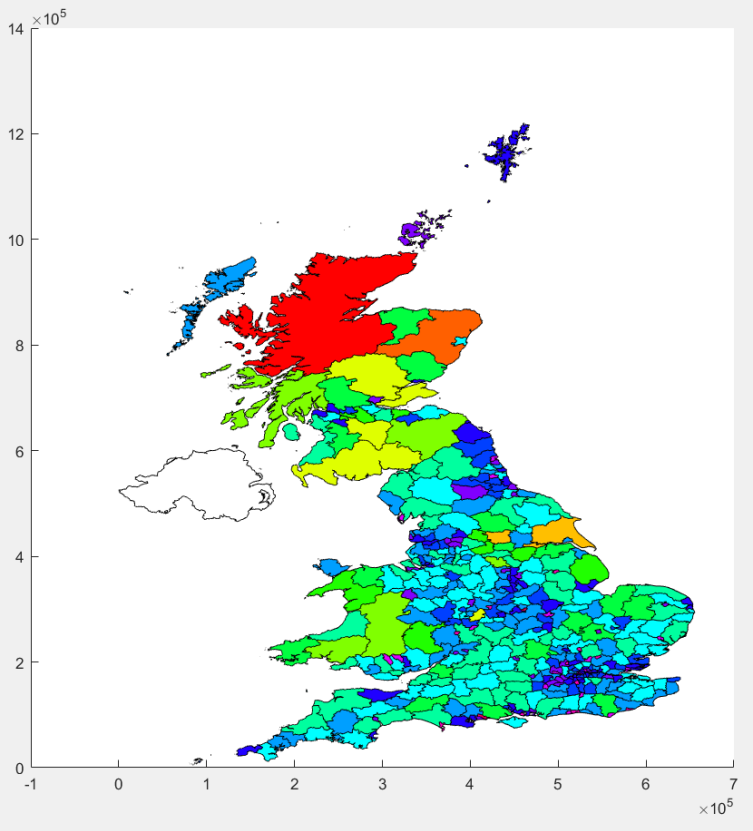


Figure 21 : Frequency distribution Log infection share week 5

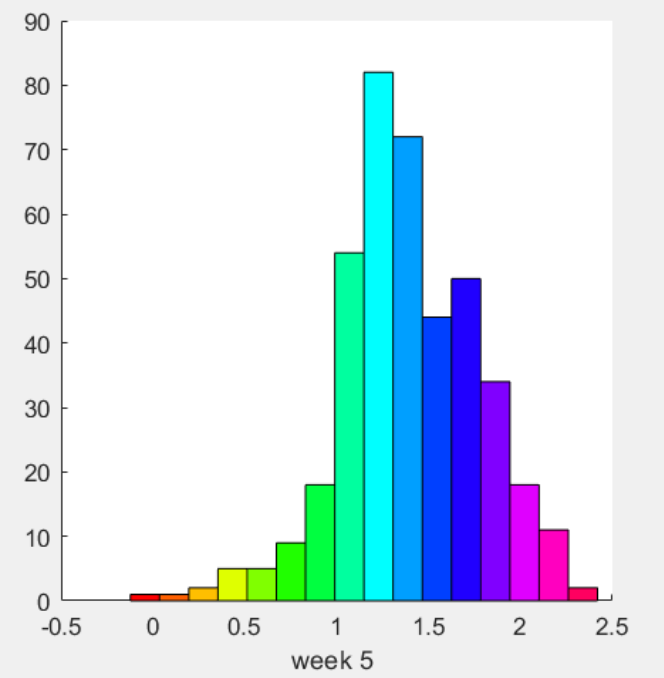


Figure 22 : Change in infection share weeks 1 to 5

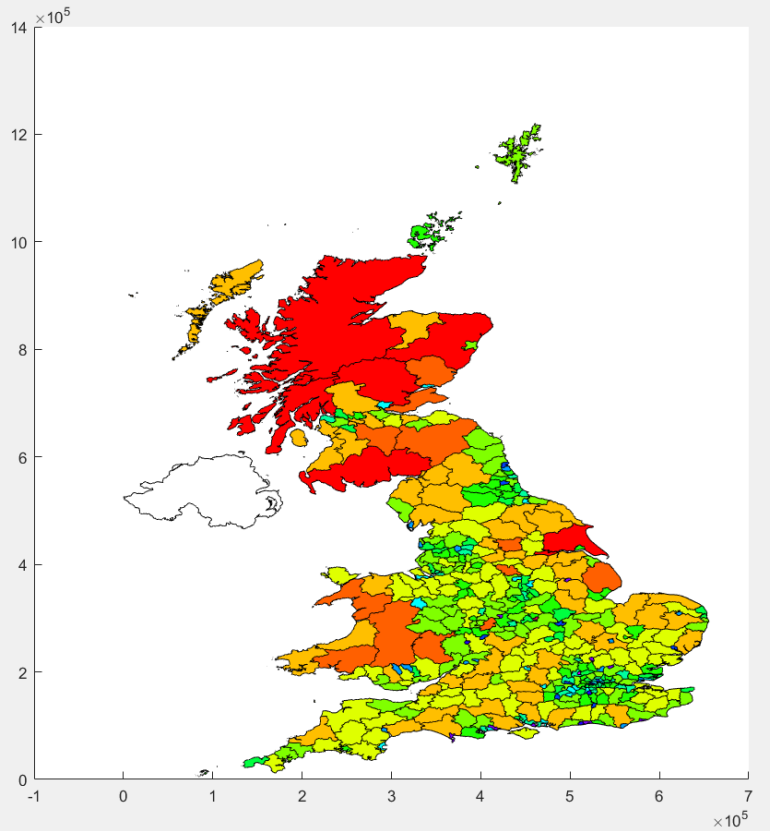


Figure 23 : Frequency distribution of change in infection share

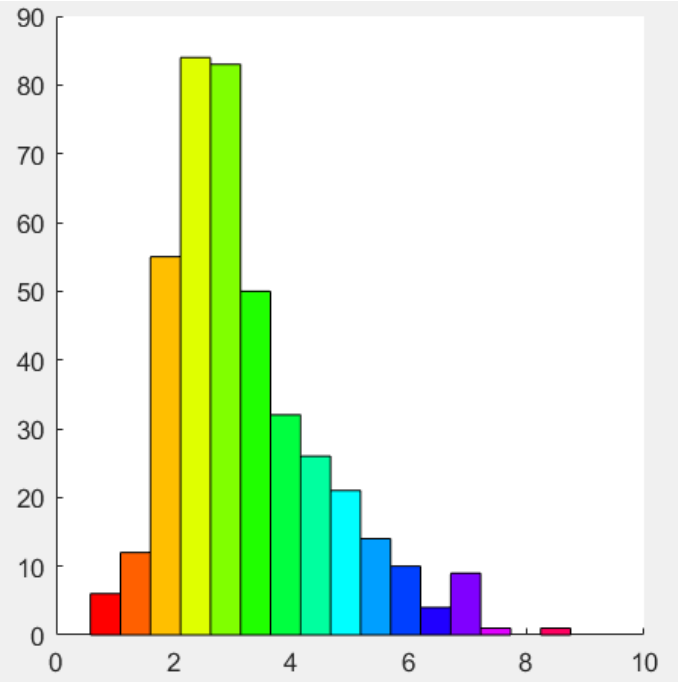
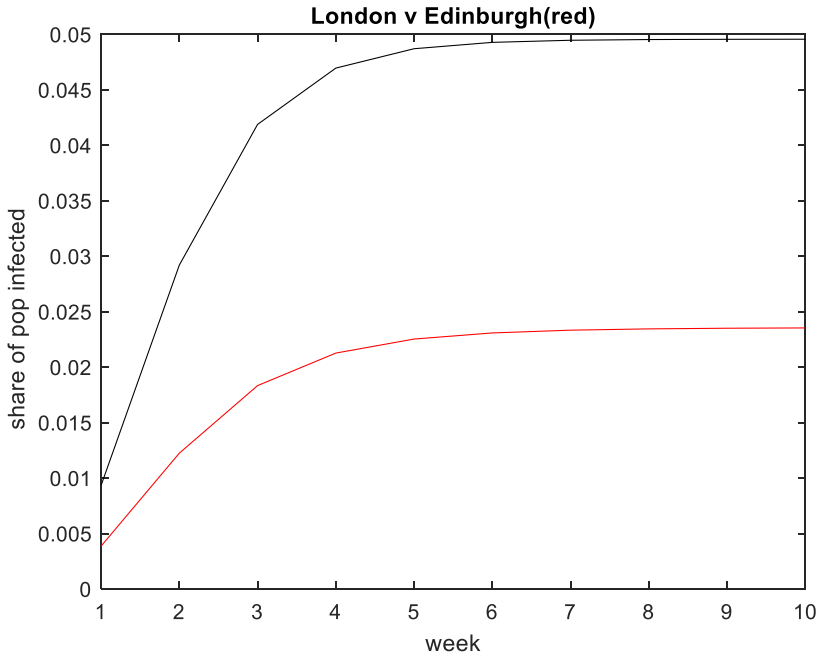


Figure 24 : infection shares in two cities



Refrences

Baltagi BH, Fingleton B, Pirotte A (2014) Estimating and forecasting with a dynamic spatial panel model. Oxford Bulletin of Economics and Statistics 76:112–138

Baltagi B, Fingleton B, Pirotte A (2019) 'A Time-Space Dynamic Panel Data Model with Spatial Moving Average Errors' Regional Science and Urban Economics 76 13-31

Fingleton B, Nikodem Szumilo(2019) Simulating the impact of transport infrastructure investment on wages: a dynamic spatial panel model approach, Regional Science and Urban Economics 75 148-164