The impact of spatial clustering on disease transmission and optimal control:

Supplementary Information

We now present some of the formal material relevant to the main text, and additional sensitivity results concerning the main conclusions.

Methodological approach for the Optimization Procedure

In order to investigate the accuracy of random location models, these models must first be parameterised to match epidemics on the truth data. This is done in the following way:

For the UK counties, we use the parameters which have previously been identified to provide the best fit to the UK 2001 FMD epidemic (Tildesley *et al* 2008) and run 1000 epidemics within each county using the 2001 farm demography truth data for those counties (Cumbria, Devon, Aberdeenshire and Clwyd). For each epidemic, on each day the cumulative number of farms reported and culled as well as the cumulative number of cattle and sheep on such farms (parameter γ) is stored. Ignoring any epidemics which do not take off, one epidemic is then selected at random for the matching procedure (we note that the particular choice of epidemic is incidental - this work is focused on the ability of random location models to capture epidemic behaviour and control that would be seen on the truth data). For each county, each farm is now allocated a random location within the county borders and the random location farm database model is now re-parameterised to provide a "best fit" to the epidemic seen on the truth data. This parameterisation is achieved by running 10000 simulations of the random location model and allowing the transmission kernel parameters (kernel width and height, K_w and K_h respectively) to vary such that the average difference on a daily basis between 'simulated epidemics' from the random location data to the 'observed epidemic' simulated on the true spatial data for parameter γ is minimised. Once this optimization has been carried out, 10000 simulations are run for the "truth data" model and for the random location model (with "best fit" parameters) with ring culling including, to determine the optimal ring cull radius for each demography which minimises the Epidemic Impact.

In the US scenario, for all four counties data sets investigated (Lancaster PA, Franklin TX, Cuming NE and the adjoining counties of Wright and Humboldt in Iowa), the same procedure is carried out. However, owing to much lower livestock densities in the US, should the UK parameters be applied directly to the US system, epidemics in these five counties are restricted to a handful of farms in each case. However, it would be naive to assume that the same transmission kernel would be valid for an outbreak of FMD in USA. For the US counties the kernel width K_w is scaled for the truth data in each county such that:

$$\frac{K_w(US)}{K_w(UK)} = \frac{D_f(UK)}{D_f(US)} \tag{1}$$

where $K_w(UK)$ is the width of the UK dispersal kernel, $D_f(UK)$ is the overall density of farms in the UK whilst $K_w(US)$ and $D_f(US)$ are the width of the dispersal kernel and the overall density of farms in the US county under consideration respectively. This scaling will, on average, preserve the overall number of contacts per farm and result in much larger epidemics. The Epidemic Impacts presented in table 1 of the main text are therefore not an indication of Epidemic Impacts we would necessarily expect in the event of an epidemic but are a feature of this scaling. The scaling itself is carried out to investigate robustness of the random location model results in the US scenario – a random model will fit very easily to epidemics which do not take off and optimal control policies for both the random model and truth data model will be identical (i.e. no ring culling); when the epidemics do take off the fitting procedure is much more complex and preferred ring culling strategies are not obvious without simulation.

Random density

In order to investigate clustering properties of the farm data discussed in the main body of the paper, we consider a population of N farms, indexed by $i = 1, \ldots, N$. Then $N_i(r)$ is the number of farms a distance r from farm i and

$$D_i(r) \coloneqq \frac{N_i(r)}{\pi r^2} , \quad N = N_i(\infty) + 1, \forall i . \tag{2}$$

We are particularly interested in the mean density

$$D(r) \coloneqq \frac{1}{N} \sum_i D_i(r) , \tag{3}$$

although we may also wish to consider measures of farm-level variability such as prediction intervals. In the case of an infinite population with random locations, we expect $D(r) = \bar{D}$. For a finite $l \times l$ square, with internal uniform density of farms $\bar{D} = N/(l^2)$ and no external farms,

$$D(r) = \bar{D}\left(1 - \frac{r}{l}\right) + O\left(\frac{r}{l}\right)^2 . \tag{4}$$

This explains the negative linear slope seen in spatially unclustered farm networks in the main text.

Generation of non-random spatial arrangement

Given a number of points in a unit square we generate a spatial arrangement with target mean distribution $D(r)$ using the following highly schematic algorithm:

1. Assign x- and y-locations randomly to each point i, giving $\mathbf{r}_i = (x_i, y_i)$.
2. Calculate the current mean distribution $\tilde{D}(r)$ and calculate the error $\tilde{\epsilon} = |D - \tilde{D}|^2$.
3. Pick a random point j and propose a new location $\hat{\mathbf{r}}_j = (x_j + \delta x, y_j + \delta y)$ where $\delta x, \delta y$ are drawn from a normal distribution of unit variance until the proposed location lies within the square.
4. Calculate the modified distribution $\hat{D}(r)$ if the location of j were changed to the proposal.
5. If $\hat{\epsilon} = |D - \hat{D}|^2 > \tilde{\epsilon}$, accept by setting $\mathbf{r}_j \to \hat{\mathbf{r}}_j$ and $\tilde{D} \to \hat{D}$. Otherwise leave the system unchanged.
6. Repeat steps 3–5 until convergence or failure.

Empirically, we find convergence for all the distributions that we wish to consider.

Parameterisation of non-random spatial arrangement

We are interested in monotonically decreasing functional forms for $D(r)$, which means from the theory of Z transforms that we can write the mean density in the form

$$D(r) = S + \sum_{m=1}^{m_{\max}} A_m e^{-k_m r} , \tag{5}$$

where $m_{\max}$ represents the number of terms used in the sum of exponentials, and the total number of parameters in the model will be $M = 1 + 2m_{\max}$. Clearly, increasing M will provide a more accurate description of the mean spatial arrangement, but potentially at the expense of introducing excess parameters into the model. We therefore make use of the Akaike information criterion (AIC, Akaike 1974) to justify an Ansatz for mean density. Given a set of n distances, at which the mean density is sampled, we use non-linear least squares fitting to fit a model to the data points $\{D(r_a)\}$ for $a = 1, \ldots, n$. For a model prediction $\{\tilde{D}(r_a)\}$ the residual sum of squares is

$$\mathcal{R} = \sum_{a=1}^{n} \left(D(r_a) - \tilde{D}(r_a)\right)^2 . \tag{6}$$

We then try to minimise the standard AIC:

$$A = 2M + n\left(\ln\left(\frac{2\pi\mathcal{R}}{n}\right) + 1\right) . \tag{7}$$

This process is applied in Supplementary Figure **??**. Pane **??** shows the distance kernel for infection, which motivates a set of distance bins and a distance cutoff. The binned mean density distribution of farms defined in (3) in Cumbria is shown in pane **??**, together with prediction intervals showing the variability at individual farm level. This binned mean is shown as a series of circles in pane **??**. A sequence of fitted curves corresponding to different values of M within the general Ansatz of equation (5) is also shown in pane **??**. The value of A as defined above for each of these is shown in pane **??**, showing that the value of $M = 3$ used in the text provides a good fit, and $M = 5$ (investigated below) is actually preferred. $M = 7$ is formally optimal – over this value extra parameters do not provide a better adjusted fit, however given both the extremely small information-theoretical likelihood gain and our epidemiological results we do not consider this extra complexity to be justified.

Results for the two-exponential model

For the two-exponential model, we again generate farm populations in a 50×50 km grid, with a variety of density profiles as defined by the parameters in equation 2 in the main text. In order to explore networks with highly spatial clustering, we assign the following ranges to the parameters in equation 2

$$\begin{aligned} 1 \leq S_0 &\leq 100 \\ 0.1 \leq S_{\text{inf}} &\leq 1.0 \\ 0 \leq A &\leq 1 \\ 0.1 \leq B_0 &\leq 10 \end{aligned} \tag{8}$$

With these parameter values, farm networks of 1000 and 2000 premises are generated. We note that, given that all other parameters are defined, in addition to the number of farms in the network and the area of the domain, the final parameter B_1 can be calculated in terms of the remaining parameter values. The same investigation as desribed in the main text is carried out, in that epidemics are simulated on the generated farm networks and a random location model is fitted to each one with variation of K_h and K_w. Each network is then investigated for the optimal value of ring cull radius which minimises the epidemic impact for the random and "truth" networks.

The results for networks of 2000 farms are summarised in Supplementary Figure 2 (the same overall behaviour is found for 1000 farm networks). As S_0/S_{inf} is increased, with A and B_0 fixed, farms become increasingly spatially clustered (panes 2a–d). For farm networks with low S_0/S_{inf} and a corresponding low degree of clustering (panes 2a,b), optimal ring cull radius is independent of precise knowledge of farm location after reparameterisation (panes 2a,b). As S_0/S_{inf} is increased, farms appear more spatially clustered (panes 2c,d) and a random location model tends to overestimate optimal ring cull radius (panes 2c,d). In the case of a 2000 farm network, the increased overall density means that there is more spatial structure apparent between clusters than for a 1000 farm network. The effect of over-culling in this case results in an increase in epidemic impact of around 20% (or around 25 farms) over ring culling at the optimal radius in the worst case scenario (pane 2c). However, ring culling at this radius still "saves" around 88% of the farms that would have livestock culled if no ring culling were carried out. We again observe a flattening out of the epidemic impact curve for the truth data in this case (pane 2c). For very highly clustered farm networks ($S_0/S_{\text{inf}} = 1000$), whilst the random model still predicts larger ring sizes than the optimal value on the truth data, the effect of this over culling is reduced (pane 2d). The effects of increasing B_0 (with $S_0/S_{\text{inf}} = 250$ and $A = 0.8$) are summarised in panes 2e–h. A similar effect to increasing S_0/S_{inf} is observed – increasing B_0 causes farms to be increasingly densely clustered (panes 2e–h). The random data again slightly overpredicts the optimal ring cull radius with, in a worst case scenario, a corresponding increase in epidemic impact of around 20% (or 40 farms, pane 2e). As farms become increasingly clustered, the effect on epidemic impact of this overculling is found to decrease (panes 2f–h). Therefore, as the density of the farm network increases, it is not, as might be naively suggested, the very highly clustered demographies

that would result in the largest increase in epidemic impact if the random model were assumed. Rather it is the "larger cluster" networks, with limited intermediate structure between clusters, for which the random model causes the greatest increase in epidemic impact. We note that of all the county data analysed in this paper, none are as clustered as these "intermediate" cases and it is important to establish whether such farm demographies exist in practice.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, **19(6)**, 716–723.