Supporting Information

Tildesley et al. 10.1073/pnas.0909047107

SI Text

Methodological Approach for the Optimization Procedure. To investigate the accuracy of random location models, these models must first be parameterized to match epidemics on the spatially clustered 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 foot-and-mouth disease (FMD) epidemic (S1) and run 1,000 epidemics within each county using the 2001 farm demography 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), are stored. Ignoring any epidemics that do not takeoff, 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 behavior and control that would be seen on the spatially clustered 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 reparameterized to provide a "best fit" to the epidemic seen on the "true" data. This parameterization is achieved by running 10,000 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 minimized. Once this optimization has been carried out, 10,000 simulations are run for the true data model and for the random location model (with best fit parameters) with ring culling included, to determine the optimal ring-cull radius for each demography which minimizes the epidemic impact.

In the US scenario, for datasets of all four counties investigated (Lancaster, PA; Franklin, TX; Cuming, NE; and the adjoining counties of Wright and Humboldt, IA), the same procedure is carried out. However, owing to much lower livestock densities in the United States, should the United Kingdom 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 the United States. For the US counties, the kernel width K_w is scaled for the true data in each county such that

$$\frac{K_w(\text{US})}{K_w(\text{UK})} = \frac{D_f(\text{UK})}{D_f(\text{US})} \qquad \text{[S1]}$$

where $K_w(\text{UK})$ is the width of the UK dispersal kernel, $D_f(\text{UK})$ is the overall density of farms in the United Kingdom, and $K_w(\text{US})$ and $D_f(\text{US})$ 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 that do not takeoff and optimal control policies for both the random model and true data

model will be identical (i.e., no ring culling); when the epidemics do takeoff, the fitting procedure is much more complex and preferred ring-culling strategies are not obvious without simulation.

Random Density. 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 $i = 1, ..., N$. Then $N_i(r)$ is the number of farms a distance r from farm i, and

$$D_i(r) := \frac{N_i(r)}{\pi r^2}, \qquad N = N_i(\infty) + 1, \ \forall i. \qquad \text{[S2]}$$

We are particularly interested in the mean density

$$D(r) := \frac{1}{N} \sum_i D_i(r), \qquad \text{[S3]}$$

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 $D(r) = D$. For a finite $l \times l$ square, with internal uniform density of farms $D = N/(l^2)$ and no external farms,

$$D(r) = D\left(1 - \frac{r}{l}\right) + O\left(\frac{r}{l}\right)^2. \qquad \text{[S4]}$$

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

Generation of Nonrandom 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 $\mathbf{r}_j \rightarrow \hat{\mathbf{r}}_j$ and $\tilde{D} \rightarrow \hat{D}$. 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.

Parameterization of Nonrandom 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

$$D(r) = S + \sum_{m=1}^{m_{\max}} A_m e^{-k_m r}, \qquad \text{[S5]}$$

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 (S2) to justify an ansatz for mean density. Given a set of n distances, at which the mean density is sampled, we use nonlinear least squares fitting to fit a model to the data points $\{D(r_a)\}$ for $a = 1, ..., n$. For a model prediction $\{\tilde{D}(r_a)\}$, the residual sum of squares is

$$\mathcal{R} = \sum_{a=1}^{n} (D(r_a) - \tilde{D}(r_a))^2. \qquad \text{[S6]}$$

We then try to minimize the standard AIC:

$$A = 2M + n\left(\ln\left(\frac{2\pi\mathcal{R}}{n}\right) + 1\right). \qquad \text{[S7]}$$

This process is applied in Fig. S1. Fig. S1A 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 Eq. S3 in Cumbria is shown in Fig. S1B, together with prediction intervals showing the variability at individual farm level. This binned mean is shown as a series of circles in Fig. S1C. A sequence of fitted curves corresponding to different values of M within the general ansatz of Eq. S5 is also shown in Fig. S1C. The value of A as defined above for each of these is shown in Fig. S1D, 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 Eq. 2 in the main text. To explore networks with highly spatial clustering, we assign the following ranges to the parameters in Eq. 2:

$$1 \le S_0 \le 100 \qquad 0.1 \le S_{inf} \le 1.0 \qquad 0 \le A \le 1$$
$$0.1 \le B_0 \le 10 \qquad\qquad\qquad\qquad\qquad \text{[S8]}$$

With these parameter values, farm networks of 1,000 and 2,000 premises are generated. We note that, given that all other parameters are defined, in addition to the number of farms in the net-

work 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 described 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 minimizes the epidemic impact for the random and spatially clustered networks.

The results for networks of 2,000 farms are summarized in Fig. S2 (the same overall behavior is found for 1,000 farm networks). As S_0/S_{inf} is increased, with A and B_0 fixed, farms become increasingly spatially clustered (Fig. S2A–D). For farm networks with low S_0/S_{inf} and a corresponding low degree of clustering (Fig. S2A and B), optimal ring-cull radius is independent of precise knowledge of farm location after reparameterization (Fig. S2A and B). As S_0/S_{inf} is increased, farms appear more spatially clustered (Fig. S2C and D) and a random location model tends to overestimate optimal ring-cull radius (Fig. S2C and D). In the case of a 2,000 farm network, the increased overall density means that there is more spatial structure apparent between clusters than for a 1,000 farm network. The effect of overculling 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 (Fig. S2C). 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 spatially clustered data in this case (Fig. S2C). For very highly clustered farm networks ($S_0/S_{inf} = 1,000$), although the random model still predicts larger ring sizes than the optimal value on the spatially clustered data, the effect of this overculling is reduced (Fig. S2D). The effects of increasing B_0 (with $S_0/S_{inf} = 250$ and $A = 0.8$) are summarized in panes Fig. S2E–H. A similar effect to increasing S_0/S_{inf} is observed: increasing B_0 causes farms to be increasingly densely clustered (Fig. S2E–H). The random data again slightly overpredict the optimal ring-cull radius with, in a worst-case scenario, a corresponding increase in epidemic impact of around 20% (or 40 farms, Fig. S2E). As farms become increasingly clustered, the effect on epidemic impact of this overculling is found to decrease (Fig. S2F–H). Therefore, as the density of the farm network increases, it is not, as might be naively suggested, the very highly clustered demographics 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 analyzed in this paper, none are as clustered as these "intermediate" cases, and it is important to establish whether such farm demographics exist in practice.

1. Tildesley MJ, et al. (2008) Accuracy of models for the 2001 foot-and-mouth epidemic. *Proc R Soc Lond B* 275(1641):1459–1468.

2. Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Automat Contr*, 19(6):716–723.
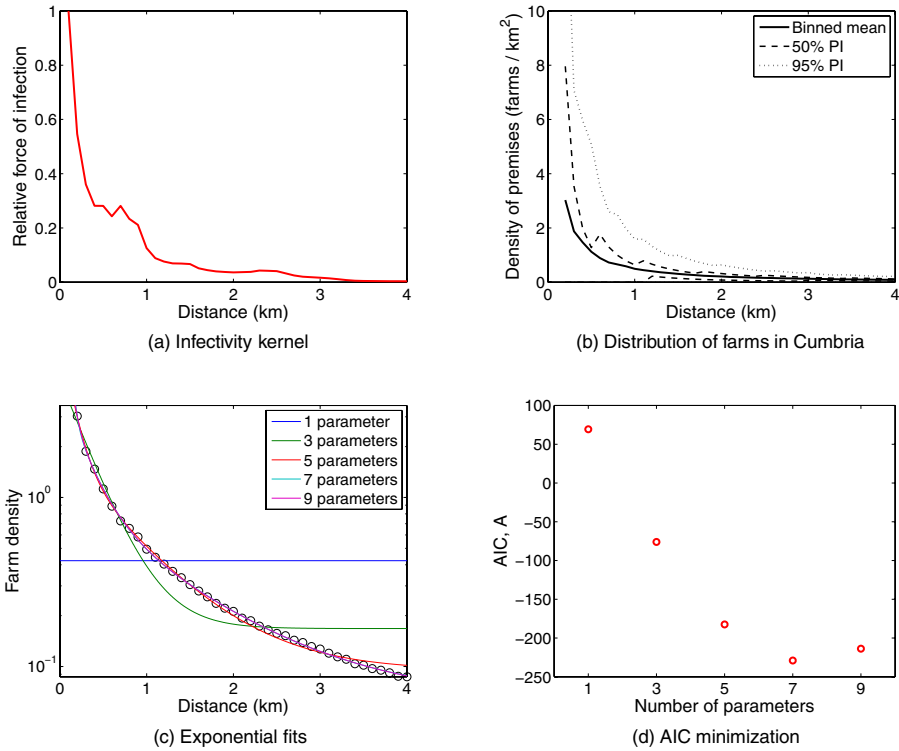
Fig. S1. Parameterization of spatial clustering in Cumbria. (A) Model infectivity kernel, which motivates a set of distance bins and a cutoff. (B) the mean binned density is shown together with the intervals within which 50% and 95% of farm-level binned density profiles sit. A sequence of exponential fits and their AIC values are shown in (C) and (D).
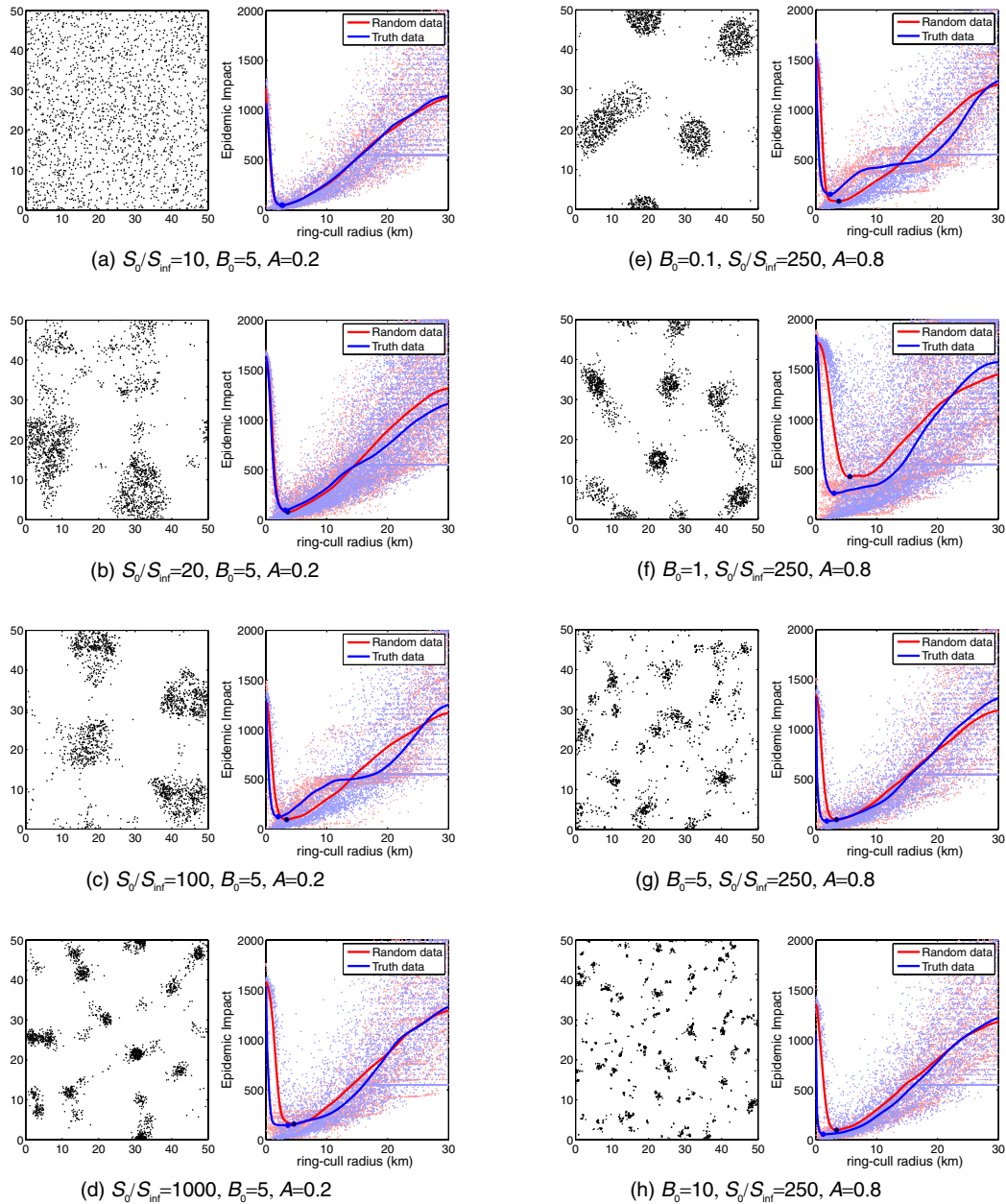
Fig. S2. Spatial location of farms (first column) and mean epidemic impact against ring cull radius for the random data (red line) and the spatially clustered data (blue line) for farm networks with different degrees of spatial clustering. Each network consists of 2,000 farms. In (A)–(D), $B_0 = 5$ and $A = 0.2$ with S_0/S_{inf} varying such that (A) $S_0/S_{inf} = 10$, (B) $S_0/S_{inf} = 20$, (C) $S_0/S_{inf} = 100$, and (D) $S_0/S_{inf} = 1,000$. In (E)–(H), $S_0/S_{inf} = 250$, $A = 0.8$ with B_0 varying such that (E) $B_0 = 0.1$, (F) $B_0 = 1$, (G) $B_0 = 5$, and (H) $B_0 = 10$. For the epidemic impact panes, raw data for both the random data (red dots) and spatially clustered data (blue dots) are also given.