Statistical modeling of holding level susceptibility to infection during the 2001 foot and mouth disease epidemic in Great Britain

Citation for published version:

Digital Object Identifier (DOI):
http://dx.doi.org/10.1016/j.ijid.2009.05.003

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published In:
International Journal of Infectious Diseases

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Statistical modeling of holding level susceptibility to infection during the 2001 foot and mouth disease epidemic in Great Britain

Paul R. Bessell a,*, Darren J. Shaw b, Nicholas J. Savilla, Mark E.J. Woolhouse a

a Centre for Infectious Diseases, University of Edinburgh, Kings Buildings, West Mains Road, Edinburgh EH9 3JT, UK
b Veterinary Clinical Sciences, R(D)SVS, The University of Edinburgh, Easter Bush Veterinary Centre, Roslin, Midlothian, Scotland

Introduction

Foot and mouth disease (FMD) is a highly infectious viral disease affecting cloven-hoofed animals.1 In the UK there was just one outbreak of FMD between 1968 and 2001.2 However, in 2001 there was a major epidemic of FMD lasting seven months with infection reported on 2026 premises (known as infected premises, IPs) in Great Britain (GB - comprising mainland England, Scotland and Wales). Within GB a further 8131 dangerous contact (DC) holdings had animals culled under different policies. These premises include ‘traditional DCs’, contiguous premises (CPs), 3 km culls–sheep premises within 3 km of an IP and slaughter on suspicion (SOS) premises–premises on which infection could not be ruled in or out on clinical grounds.3,4 With the exception of SOS premises, a number of these non-IPs may have had animals latently infected with FMD, because in many instances freedom from infection was never established through laboratory testing.4

Descriptions of the 2001 UK FMD outbreak have been extensively published elsewhere3–5 but of importance here is that the early stages of the 2001 FMD outbreak were characterized by widespread seeding of infection throughout the country by animal movements.5,7 A national movement ban (NMB) on susceptible species was implemented on 23 February 2001, after which disease transmission became spatially localized with 80% within 10 km.9 As a result, FMD remained largely restricted to the areas in which it was seeded.10

During the period of local spread, the distance to a source of infection appeared to be the principal risk factor for infection.4,10 However, despite the wealth of literature on the 2001 epidemic, potential risk factors over and above the distance to a source of infection have not been investigated explicitly. Such factors might include the species of animals on the holding, the numbers or densities of animals in the locality, and the area of the holding.

The dynamics of the epidemic during the period of local spread have been mathematically modeled using spatially explicit stochastic10,11 and deterministic12,13 mathematical models. The models of Keeling et al.10 and Tildesley et al.11 are spatially explicit...
simulations in which the next generation of infections is dependant upon the spatial location of the present generation. The predictive accuracy of the model of Tildesley et al. has been evaluated by comparing the predicted IPs with the actual IPs; the resulting accuracy is between 5% and 15%. Given the small number of IPs relative to the number of farm holdings, this represents good accuracy. However, no equivalent statistical evaluation with the potential to directly interrogate the data has been carried out on the dynamics of the epidemic during the period of local spread. Furthermore, there have been no studies that have explicitly attempted to identify risk factors for FMD infection.

This paper, therefore, will develop a statistical model of holding level risk of susceptibility to FMD. By incorporating a range of risk factors in a multivariate model, more subtleties of the epidemic can be determined. The ability of the model to identify IPs will then be compared to the model of Tildesley et al. This will be a means of quantifying the ability of this framework to discriminate between IPs and non-IPS in comparison to previously employed frameworks.

**Materials and methods**

**Data**

Farm-level data for the 2001 epidemic were taken from the June 2000 agricultural censuses conducted by the Ministry of Agriculture, Fisheries and Food (MAFF–Department for Environment Food and Rural Affairs (DEFRA)) from June 2001 in England and Wales and the Scottish Executive Environment and Rural Affairs Department (SEERAD) in Scotland. This is a register of all farm holdings and records the coordinates, farm area, and numbers of animals by species for all farm enterprises under a County-Parish-Holding (CPH) identifier. The agricultural census records the main farm unit (buildings and fields) and all associated off-fields (fields that are not part of the main farm unit) under a single CPH number. A full census was carried out in June of 2000 and has been used in many previous studies of the 2001 FMD epidemic.

Data relating to the 2001 FMD epidemic were recorded by MAFF/DEFRA in the Disease Control System (DCS) database. The DCS records data on the epidemiology of the disease on each IP, including estimated dates of infection and dates of reporting and slaughter. The period of predominantly local spread is analyzed here, so only IPs estimated to have been infected after the NMB were used as the outcome variable; this reduced the number of IPs to 1948. IPs with only pigs were excluded from the model as pig premises played little part in the epidemic after the index case; after the NMB, four pig-only premises (excluded) and 58 premises with pigs and sheep or cattle (included) became IPs.

The DCS data comprised point locations and a CPH number for all culled premises. In the DCS data, fields containing animals culled as part of the FMD control policy were recorded under a separate CPH number if the field was more than 1 km from the main farm holding; these parcels of land were therefore treated as separate holdings for the purposes of epidemic management. However, these parcels of land do not appear in the agricultural census, which means that for any analysis of risk of susceptibility to infection (which looks at the demographic details of all farms—not just those involved in the 2001 outbreak), 444 IPs (22%) are not in the agricultural census and therefore the denominator. Therefore, only the remaining 1527 premises that appear in both the FMD data as IPs and the agricultural census under the same CPH number were used as the outcome variable. Out of 8820 holdings containing animals culled as non-IPS, 4099 were on the census, the remaining 4721 were parcels of land recorded under a separate CPH number to its main holding.

There are 131 663 premises listed on the June 2000 agricultural census that had sheep or cattle registered on them. Not all of these holdings were exposed to infection during the epidemic, therefore only the 83 124 holdings that fall in one of the 37 counties in which there was an IP were included in the analysis. The county of the holding is identified by the CPH number.

**The model**

The model is a logistic regression model in which the data points are all premises on the agricultural census with cattle or sheep recorded. The outcome variable was whether a holding was one of 1527 IPs (cases) or one of 80 070 non-IPS (controls, and the reference level). The model construction follows the methodology laid out in Hosmer and Lemeshow.

The predictor variables fall into the four categories below:

1. **Distance to a seed IP**: The Euclidean distance in km to the nearest of 78 seed IPs, which were estimated to have been infected before the NMB. These seed IPs are identified using the DCS data on estimated dates of infection.
2. **Regional data–county of the farm**: There were 37 counties involved in the study, of which 27 had fewer than 50 IPs. These 27 counties were aggregated to test as a predictor, which was the reference level for the 11 level factor (Figure 1).
3. **Holding data**: Livestock species on the holding (mixed, cattle only, or sheep only, with mixed the reference level) and area (in Ha) of the holding. Species presence rather than animal numbers on the holding was used because there are a large number of holdings with sheep and no cattle (24.6%) and cattle and no sheep (36.1%), therefore, this variable could not be treated as a continuous distribution.
4. **The neighborhood of the farm**: Total numbers of cattle and total numbers of sheep within 10 km of the holding.

Control effort factors including biosecurity, surveillance, and pre-emptive culling have not been explicitly included in the model. However, the control effort varied between counties, which are accounted for here.

Predictors were analyzed in univariate models to screen for risk factors to be entered into subsequent multivariate modeling. Distance to seed and total cattle and sheep in the locality were square root transformed and holding area was log$_{10}$ + 1 transformed to linearize the logits. Predictors from the univariate models that were significant at $p < 0.25$ were included in the multivariate model. Variables were analyzed in the multivariate model and those that were significant at $p < 0.05$ were retained in the model and biologically plausible interactions were tested. Changes to the estimates of the variables were monitored and new variables were added as this could indicate intercorrelation. This was further tested by dropping terms and inspecting the influence of remaining variables.

The discriminatory power of the model was evaluated by taking the 1527 greatest fitted values as predicted IPs. During the model building, AIC was used to compare between models in which the outcome variable was the same. The percentage of these predicted IPs that was actually IPs gave a measure of the ability of the model to discriminate between IPs and non-IPS. The discriminatory power of the model was further analyzed by plotting the receiver operator characteristic (ROC) curve for the model. The ROC curve plots the sensitivity and specificity of the model given a range of cut-off values, where the cut-off value is the threshold that determines whether a test result is considered positive or negative. The sensitivity is the proportion of IPs that would be considered positive given its fitted value falling above a defined cut-off, and the specificity is the proportion of non-IPS that test
negative given their fitted value falling below that cut-off value. The cut-offs used to generate this ROC curve are 1000 evenly spaced points through the range of fitted values. The goodness of fit of the model was evaluated by calculating the area under the ROC curve (described above) and the Hosmer and Lemeshow Chi-square statistic, both of which use rankings of modeled values to give a measure of the overall model fit (or lack of fit in the case of the Hosmer and Lemeshow test).18

Spatial autocorrelation in the model was evaluated by carrying out a Moran’s I test for spatial autocorrelation on the model residuals.20 Spatial autocorrelation was evaluated across a range of numbers of nearest neighbors to test for significant \( p < 0.05 \) autocorrelation at a range of scales. Significant autocorrelation was found at all scales (Moran’s I statistic \( p < 0.001 \)) and was adjusted for by overlaying a lattice of 5 km wide (edge–edge) hexagons generated using the repeating shapes21 extension for ESRI ArcView 3.2 onto the farm distribution. Similar approaches have been used elsewhere to overcome this problem,22,23 however these studies used political boundaries. For this study, a hexagonal grid was used in preference to political boundaries such as parishes, as the hexagon boundaries are arbitrary and all polygons are the same size. The grid of hexagons that covered the susceptible population in this study comprised 5033 hexagons, each one 21.6 km\(^2\), with a median of 14 farms in each (25th percentile = 6, 75th = 24). The unique ID of the hexagon in question was incorporated into the model as a random effect to form a generalized linear mixed model (GLMM) with binomial errors. The residuals of the GLMM were tested for spatial autocorrelation using the previously described method.

All modeling was carried out in the R Statistical environment24 using the spdep package to calculate Moran’s I and the nmle4 package to generate the GLMMs.

**Results**

**Univariate analysis**

In univariate analysis, all predictors were statistically significant at the \( p < 0.001 \) level (Table 1). The principal predictors were the county of the farm, with farms in Cumbria at the greatest risk of infection. Mixed farms were at significantly greater risk than cattle farms, which in turn were at greater risk than sheep-only farms. The surface area of the holding and surrounding cattle and sheep densities were also positively associated with being an IP.

**Multivariate analysis**

All predictors were entered into the multivariate GLMM (Table 2) and all were retained with the exception of sheep density, which was non-significant \( p > 0.1 \) at all scales. The fit of
the model was good; the area under the ROC is 0.962, Hosmer and Lemeshow Chi-square = 49.98 ($p > 0.1$), and it correctly identifies 688 (45.1%, 95% confidence interval (CI) = 42.6, 47.6) IPs. The ROC curve for this model (Figure 2) demonstrates that this model is very good at discriminating between IPs and non-IPs. Perfect discrimination would be demonstrated by a cut-off that maximizes both sensitivity and specificity so that the curve approaches the top left corner of the ROC curve. Furthermore, there was no spatial autocorrelation in the residuals of the GLMM (Moran's I statistic $p$-value $> 0.5$) and there was little evidence for intercorrelation between predictors.

Farms in the county of Cumbria (odds ratio (OR) = 25.38, 95% CI = 16.92, 38.07) were at the greatest risk of infection relative to farms in the ‘rest of GB’ category. Decreased distance to the infectious seed (OR = 0.613 per square root km from an infectious seed, 95% CI = 0.558, 0.672) and higher local cattle densities (OR = 3.41 (2.99, 3.89)) remained major risk factors. However, additional variables relating both to a larger holding area and the surrounding cattle density are also risk factors. The high quality of the data enabled the identification of individual seeds of infection. In a similar model constructed for Vietnam,23 the authors break the epidemic down into time periods and regions to describe the introduction of the virus into different areas in different years. For vector-borne diseases such as sleeping sickness, the point source can be modeled as the vector habitat.25 In the 2001 FMD epidemic, the location

**Discussions**

A statistical model was developed to calculate the risk of a holding on the agricultural census becoming an IP. The model was a good fit and was able to identify 45% of the actual IPs; the corresponding figure from mathematical models that simulate the 2001 FMD outbreak is between 5% and 15%.14 The NMB greatly reduced the range of disease transmission and FMD largely remained in the areas in which it was initially seeded, therefore one of the principal predictors of susceptibility to FMD is the distance to a seed infected before the NMB. The presence of FMD in the locality is further described by the county of the holding. However, additional variables relating both to a larger holding area and the surrounding cattle density are also risk factors.

Table 1

Univariate logistic regression analysis of IPs and non-IPs. Non-IPs are the reference level. The non-IP and IP columns relate to the number of each group falling into the different levels of each factor.

<table>
<thead>
<tr>
<th>Predictor Unit</th>
<th>Non-IP</th>
<th>IP</th>
<th>z-Score</th>
<th>p-Value</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance to seed</td>
<td>$\sqrt{km}$</td>
<td>80 070</td>
<td>1527</td>
<td>-32.67</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>County</td>
<td>Rest of GB</td>
<td>44 124</td>
<td>149</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Cumbria</td>
<td>4320</td>
<td>725</td>
<td>42.76</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Devon</td>
<td>7908</td>
<td>135</td>
<td>13.57</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Dumfrieshire</td>
<td>961</td>
<td>108</td>
<td>26.86</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Durham</td>
<td>1498</td>
<td>73</td>
<td>18.38</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Gloucestershire</td>
<td>2181</td>
<td>51</td>
<td>11.82</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Hereford and Worcester</td>
<td>4179</td>
<td>41</td>
<td>6.02</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Lancashire</td>
<td>3153</td>
<td>37</td>
<td>6.75</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Northumberland</td>
<td>1744</td>
<td>66</td>
<td>16.13</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>N. Yorkshire</td>
<td>5599</td>
<td>99</td>
<td>12.75</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Powys</td>
<td>4443</td>
<td>43</td>
<td>6.06</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Species</td>
<td>Mixed</td>
<td>30 153</td>
<td>996</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Cattle</td>
<td>29 584</td>
<td>408</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Sheep</td>
<td>20 333</td>
<td>123</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Holding area log$_{10}$ Ha + 1</td>
<td>80 070</td>
<td>1527</td>
<td>25.26</td>
<td>&lt;0.001</td>
<td>14.43 (10.86, 19.18)</td>
</tr>
<tr>
<td>Cattle density (kernel transformed) $\sqrt{10}$ head</td>
<td>80 070</td>
<td>1527</td>
<td>18.34</td>
<td>&lt;0.001</td>
<td>5.27 (4.08, 6.81)</td>
</tr>
<tr>
<td>Sheep density (kernel transformed) $\sqrt{10}$ head</td>
<td>80 070</td>
<td>1527</td>
<td>18.34</td>
<td>&lt;0.001</td>
<td>5.27 (4.08, 6.81)</td>
</tr>
</tbody>
</table>

Table 2

Multivariate GLMM analysis of risk factors for FMD infection during the 2001 epidemic in GB.

<table>
<thead>
<tr>
<th>Predictor Unit</th>
<th>z-Score</th>
<th>p-Value</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>–20.73</td>
<td>&lt;0.001</td>
<td>0.613 (0.558, 0.672)</td>
</tr>
<tr>
<td>Distance to seed $\sqrt{km}$</td>
<td>–10.31</td>
<td>&lt;0.001</td>
<td>25.38 (16.92, 38.07)</td>
</tr>
<tr>
<td>County</td>
<td>Rest of GB</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Cumbria</td>
<td>15.63</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Devon</td>
<td>6.970</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Dumfriesshire</td>
<td>6.729</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Durham</td>
<td>6.203</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Gloucestershire</td>
<td>2.673</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Hereford and Worcester</td>
<td>0.918</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Lancashire</td>
<td>0.918</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Northumberland</td>
<td>0.918</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>N. Yorkshire</td>
<td>0.918</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Powys</td>
<td>0.918</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Species</td>
<td>Mixed</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Cattle</td>
<td>–5.807</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Sheep</td>
<td>–4.820</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Holding area log$_{10}$ Ha + 1</td>
<td>18.34</td>
<td>&lt;0.001</td>
<td>4.493 (3.826, 5.275)</td>
</tr>
<tr>
<td>Cattle density (kernel transformed) $\sqrt{10}$ head</td>
<td>8.690</td>
<td>&lt;0.001</td>
<td>5.211 (4.025, 6.772)</td>
</tr>
</tbody>
</table>

GLMM, generalized linear mixed model; FMD, foot and mouth disease; OR, odds ratio; CI, confidence interval.
of disease introductions was known precisely, and these were included as point sources for infection. However, in the absence of a movement ban, the range of spread would be over much greater distances and the nature of the effect of distance to an infectious seed would be significantly less, particularly as the epidemic progressed. Previous studies have demonstrated the effect of the NMB on controlling longer range spreads of FMD.

In addition to data regarding the point sources of virus introduction, the nature of the susceptible holding is also important in determining its susceptibility. Mixed cattle and sheep holdings are at statistically significantly greater risk than single species holdings; farms covering a greater surface area are also at greater risk because they will generally have more animals, although this varies with the type of farming. Furthermore, larger numbers of cattle in a 10-km radius surrounding a holding is an important risk factor. In spite of the importance of higher cattle densities in this epidemic, sheep density was non-significant in multivariate analysis. This may reflect the infectious challenge to the holding–sheep have been estimated to be an order of magnitude less infectious than cattle. Whilst this analysis has included demographic risk factors over the entire population of holdings–sheep have been estimated to be an order of magnitude more susceptible than cattle, in spite of the type of farming. Furthermore, larger numbers of cattle in a 10-km radius surrounding a holding is an important risk factor. In spite of the importance of higher cattle densities in this epidemic, sheep density was non-significant in multivariate analysis. This may reflect the infectious challenge to the holding–sheep have been estimated to be an order of magnitude less infectious than cattle, whilst this analysis has identified demographic risk factors over the entire population of susceptible farms, there remain risk factors at the much smaller scale, such as river and rail barriers, which continue to influence the pattern of FMD cases.

The model also shows that the epidemic was operating differently in different parts of the UK, over and above the contribution of other factors in the model; farms in Cumbria were at greatest risk of infection, whilst those in Devon were at relatively low risk compared to some counties with fewer IPs such as Dumfriesshire and Gloucestershire (Tables 1 and 2). These differences may reflect regional heterogeneities in the nature of farming or farming practices, such as the distribution of animals in fields or biosecurity, which result in holdings in certain areas being more susceptible to infection with FMD. These differences may also reflect differences in the management of the epidemic by the different regional disease control centers (DCCs). In spite of the county predictor being highly significant, its removal from the model in sensitivity analysis resulted in slightly greater discriminatory power as measured by area under the ROC curve (0.976 opposed to 0.962). This is because upon the removal of county, some of that variation is taken up by other continuous predictors (in particular distance to seed); this introduces greater heterogeneity into the fitted values upon which the area under the ROC curve is based.

One of the limitations of using the GLMM framework was that the numbers of animals on the holding could not be included as a predictor due to their zero-inflated distribution. Previous analyses found a positive relationship between the numbers of animals by species on the holding and susceptibility, and have found cattle to be an order of magnitude more susceptible than sheep. This has been overcome to an extent in this study by using the holding area and the numbers of animals by species on the holding. There are regional differences in stocking densities, however broadly the holding area will be correlated with the number of animals on the premises.

In conclusion, this analysis demonstrates that post-NMB risk of FMD can be modeled relatively easily using a simple statistical framework and this approach is better at identifying the actual IPs than similar analysis using mathematical model simulations of the outbreak.

Conflict of interest
No conflict of interest to declare.

Acknowledgements

PRB is grateful to the BBSRC for providing funding. DJS, NJS and MEJW are grateful to The Wellcome Trust for support. We are also grateful to Miles Thomas from the Central Science Laboratory, DEFRA, Sand Hutton, Yorkshire for his help with data. We are also grateful to three anonymous reviewers for their valuable comments on this manuscript.

References


Figure 2. Receiver operator characteristic (ROC) curve for the multivariate model in Table 2. Sensitivity refers to the proportion of infected premises (IPs) that would be considered positive given a cut-off, specificity is the proportion of non-IPs that test negative at that cut-off.