Stochastic modeling of animal epidemics using data collected over three different spatial scales

Chris Rorres *, Sky T.K. Pelletier, Gary Smith

School of Veterinary Medicine, University of Pennsylvania, 382 West Street Road, Kennett Square, PA 19348, USA

A R T I C L E   I N F O
Article history:
Received 8 November 2010
Revised 3 February 2011
Accepted 10 February 2011
Available online 18 February 2011

Keywords:
Avian influenza
Parameter estimation
Mathematical models
ZIP-codes

A B S T R A C T
A stochastic, spatial, discrete-time, SEIR model of avian influenza epidemics among poultry farms in Pennsylvania is formulated. Using three different spatial scales wherein all the birds within a single farm, ZIP code, or county are clustered into a single point, we obtain three different views of the epidemics. For each spatial scale, two parameters within the viral-transmission kernel of the model are estimated using simulated epidemic data. We show that simulated epidemics modeled using data collected on the farm and ZIP-code levels behave similar to the actual underlying epidemics, but this is not true using data collected on the county level. Such analyses of data collected on different spatial scales are useful in formulating intervention strategies to control an ongoing epidemic (e.g., vaccination schedules and culling policies).

© 2011 Elsevier B.V. All rights reserved.

Introduction

Increases in computing power within the last decade have permitted the simulation of viral animal epidemics using stochastic, discrete-time, spatial models involving tens of thousands of farms. However, to relate these simulations to actual farm communities, accurate data are required as to the locations of the farms in the community and the numbers of animals that each farm contains. While such data are available in many countries, they are severely lacking in the United States because of confidentiality and privacy restrictions.

This study examines avian influenza (AVI) epidemics within the state of Pennsylvania. We take Pennsylvania as the spatial extent of our study because of its large farmed poultry population and because our research facilities are located in this state. Lacking accurate data as to the locations and sizes of the poultry farms in Pennsylvania, we used synthetic geographical data generated by the Research Triangle Institute (RTI) of North Carolina (Bruhn et al., 2007). RTI's synthetic geographical data provide an accurate representation of the poultry farms throughout the United States and are the geographical data we applied in the analysis of an actual AVI epidemic in 1983–84 in Pennsylvania (Buisch et al., 1984) that was useful in our present study.

The mathematical model we employ is a stochastic, discrete-time, spatial model of the type extensively in use since the 2001 foot-and-mouth epidemic in Great Britain. This type of model has been studied in detail by Keeling et al. (2001, 2003), Keeling (2005), Keeling and Eames (2005), Tildesley et al. (2006) and Tildelsley and Keeling (2008). We modified such a model to include a transmission kernel, described in the next section, which contains two parameters that can be tuned to a particular viral epidemic. These parameters determine the way in which the probability that an infectious farm will infect a susceptible farm is dependent on the distance between the two farms. In addition to the AVI epidemics examined in this paper, our transmission kernel can model animal epidemics due to foot-and-mouth disease (FMD), infectious salmon anemia (ISA), Exotic Newcastle disease (END), and other viral animal diseases.

Our choice of model is somewhat dictated by John von Neumann’s advice that the justification of a mathematical model is “solely and precisely that it is expected to work” (von Neumann, 1955). Many quite elaborate models of animal epidemics have been described and employed which require a large amount of data to implement; e.g., the North American Animal Disease Spread Model (Harvey et al., 2007). However, our experience has been that such data are extremely difficult to find. Hence, our model contains only two parameters within it, and we have resisted the temptation to include more, being reminded of another remark of von Neumann that “With four parameters I can fit an elephant, and with five I can make him wiggle his trunk” (Dyson, 2004). Generally speaking, the more parameters a mathematical model contains the less reliable it is for predictive purposes. Von Neumann’s point was that for a model with many parameters it is easy to find values of the parameters that fit given data, but it is unlikely that these parameters will fit future data.
Our main objective for modeling animal epidemics is to develop strategies for the prevention and control of such epidemics. This requires models that can correctly predict the future progress of an ongoing epidemic or accurately describe possible new epidemics.

The effect of the spatial scale used in our model on the determination of the two model parameters is of particular interest to us. While it is rare to find information as to the locations and sizes of individual farms in most states, all states provide information about the numbers of farm animals within individual counties and some also provide such data for individual ZIP codes. Our study is thus concentrated on the ability of our model to accurately describe epidemics when data are available on the farm, ZIP-code, or county levels. We thus first simulate control epidemics on the farm level and examine the consequences of having only data as to which ZIP codes or counties have been infected during the course of the epidemic, rather than which individual farms were infected.

Our model

AI epidemics are typically modeled as SEIR epidemics in which each bird is either never infected or passes through four successive stages: Susceptible, Exposed (or latent), Infectious, and Recovered. In the latent stage a bird is asymptomatic and not infectious. In the recovered stage a bird has either physically recovered from the disease, died from it, or been culled. In any case, once recovered a bird cannot become susceptible again. In our model we lump all of the birds in each of our spatial units (farm, ZIP code, or county) into a single point whose size is the number of individual birds in the spatial unit. The location of the point on the farm level is given by the farm coordinates as provided by Bruhn et al. (2007); for the ZIP-code level it is the centroid of each ZIP-code region; and for the county level it is the centroid of each county. The centroids for the ZIP codes and counties were determined using ArcGIS™, a geographical information systems software package.

The durations of the latent and infectious stages are dependent on the spatial unit used. The epidemic within a spatial unit is a small-scale version of the entire epidemic, so the durations that we assign in our model to the latent and infectious stages represent discrete approximations to intervals during which first the exposed birds dominate the spatial unit and then the infectious birds dominate the spatial unit. Consequently, the durations of these stages increase as the size of the underlying spatial units increases.

A key assumption of our model is that the probability that an infectious bird in one unit will infect a susceptible bird in another unit is the product of the probabilities that one infectious bird will infect one susceptible unit during that time step. Suppose the ith unit is infectious and contains Ni birds and the jth unit is susceptible and contains Nj birds. Then if dij is the distance between the two units, the probability that no infectious bird in the ith unit will infect any bird in the jth unit is given by

\[ p_d(\delta, \rho) = \exp\left(-N_iN_j \frac{\delta}{d_{ij}}\right). \]  

This result assumes that infectious birds infect susceptible birds independently of each other.

Finally we take into account all units that are infectious in a particular time step and determine the probability that none of the birds in the infectious units will infect any of the birds in a specified susceptible unit during that time step. If we let \( A(k) \) be the set of indices of those units that are infectious at the beginning of the kth time step of the epidemic, then

\[ p_k(\delta, \rho) = 1 - \prod_{i \in A(k)} \exp\left(-N_iN_j \frac{\delta}{d_{ij}}\right). \]  

This result is plotted in Fig. 1. The probability \( p(d) \) vs. the distance \( d \) that one infectious bird will infect one susceptible bird in one time unit for various values of the epidemic parameter \( \rho \).
On the farm level we ran our model using a 1-day time step and took the latent period to be 5 days based on data from the 1983–84 Pennsylvania AVI epidemic (Buisch et al., 1984). As to the duration of the infectious period, we considered the situation in which farms are culled once they are determined to be infectious, and 7 days was the average time from detection to culling at the farm level in the 1983–84 AVI epidemic in Pennsylvania (Buisch et al., 1984).

On the ZIP-code level and county levels we experimented with different time steps and different latent and infectious periods to find appropriate values. Because our model calls for each spatial unit to be in the same SEIR state over each time step, we considered the latent period to be that period when a sufficiently large fraction of the farms in a typical spatial unit have been infected. Accordingly, we examined the growth and decay of the number of farms in the latent state in a typical ZIP code or county over the 1000 control epidemics and decided that a 10-day latent period was appropriate for both the ZIP-code and county levels. Similarly analyzing the growth and decay of the number of farms in the infectious stage, we chose a 15-day infectious period on the ZIP-code level and a 20-day infectious period on the county level.

While a time-step of 1 day is appropriate on the farm level, it is too short for the ZIP-code and county levels. One-day time steps for the ZIP and county levels resulted in epidemics that spread too quickly. We chose time steps that yielded epidemics generated through our model that peaked at roughly the same time as our 1000 control epidemics that were generated on the farm level. We also chose time steps on the ZIP-code and county levels that were commensurate with the latent and infectious period on the corresponding level. In particular, a 5-day time step on the ZIP-code level and a 10-day time step on the county level resulted in epidemics whose gross behavior matched the behavior of the 1000 control epidemics fairly well.
Geographical data

The geographical information needed to run our simulations is displayed in Fig. 2. This synthetic map shows 7043 farms in Pennsylvania as determined by the Research Triangle Institute (Bruhn et al., 2007). RTI also generated estimates for the number of birds on each of the farms, the total number of birds for the state being about 56 million with an average of 7944 birds per farm. For our control simulations we took one of these farms, as indicated by the large disk, as the index farm: the farm that is initially infected and begins the epidemic. This farm is a fairly large one located in the southeastern portion of the state where there is a large concentration of poultry farms.

The left map in Fig. 3 displays the 805 ZIP codes in Pennsylvania that contain any of the 7043 farms, an average of 69,499 birds per ZIP code. The ZIP code that contains the index farm is shown in black. There are a total of 2190 ZIP codes in Pennsylvania, but most of them are quite small, metropolitan ZIP codes that contain no poultry farms (e.g., Philadelphia and Pittsburgh). They are located in the white regions in Fig. 3 and we have not drawn their boundaries. In the center map in Fig. 3 we have drawn a disk at the centroid of each of the 805 ZIP codes and shown the 7043 farms in gray in the background. In the right map we show only the 805 centroid disks with the index ZIP-code centroid drawn as a white disk and the other centroids drawn

![Image](image-url)
with three different radii to indicate an arbitrary classification of the number of birds in each ZIP code as small, medium, or large.

Fig. 4 shows the 67 counties of Pennsylvania, all of which contain at least one poultry farm, with the county containing the index farm in black. There is an average of 835,030 birds per county. The center map displays the farms as small gray dots and the centroids of the 67 counties are shown as black disks of three different radii. The right map shows only the 67 centroid disks with the index county as a white circle.

The 67 counties are not too unusual in shape and the distance between any two county centroids can be regarded as a reasonable measure of what we might consider as the “distance” between the two corresponding counties, and hence the distances $d_{ij}$ that we would use in our formulas. However, some of the ZIP codes in Pennsylvania have rather unusual shapes. Fig. 5 shows the detail of a few of the ZIP codes showing how, in some cases, the centroid of a ZIP code lies outside of the ZIP code and how, in some cases, one ZIP code is completely contained within another. Nevertheless, since there are fewer than 9 farms on average in each zip code, the network of Zip-code centroids seems to give a good approximation to the actual farm network.

Notice how the respective numbers of the three spatial units differ by about one order of magnitude: specifically, from 7043 farms to 805 ZIP codes we have about one-ninth as many spatial units, and from 805 ZIP codes to 67 counties we have about one-twelfth as many spatial units. The distribution of the sizes of the spatial units, in terms of number of birds each contains, is quite unequal at each spatial level. There are very many units with few birds and very few units with many birds. To illustrate: the largest farm contains as many birds as the 6055 smallest farms; the largest ZIP code contains as many birds as the 681 smallest ZIP codes; the largest county contains as many birds as the 61 smallest counties; and the largest 3 counties contain more birds than the rest of the state combined.

Our epidemic data

To see the effect of these different spatial scales on our epidemic simulations, we first chose epidemic parameters of $\delta = 5.8 \times 10^{-5}$ and $\rho = 2.50$ that we derived from our Method 2 of a previous paper (Rorres et al., 2010) using data from the 1983–84 AVI epidemic in Pennsylvania (Buisch et al., 1984). We then ran 1000 control simulations on the farm level beginning with the index farm shown in Fig. 2. The top parts of Fig. 6 and Fig. 7 display various aspects of the resulting farm-level simulations. In Fig. 6, the top left graph shows the bird attack rates of the 1000 simulations. By the bird attack rate of an epidemic we mean the fraction of birds in Pennsylvania (including the birds of the index farm) that were infected during the epidemic. The

![Fig. 7. Attack rates of the fraction of units (farms, ZIP codes, or counties) infected over 1000 simulated control epidemics.](image-url)
Fig. 8. Plots of the numbers of newly infectious birds (left) or units (right) for the three levels (farm, ZIP code, or county) as a function of time. These values were averages over the severe epidemics out of 1000 control simulations.

Fig. 9. Each of the 7043 poultry farms is shaded according to the number of times that it was infected in the severe control epidemics among the 1000 simulated.
top-left graph displays these 1000 attack rates along the vertical axis sorted along the horizontal axis according to magnitude.

From the 1000 simulated epidemics run on the farm level, we then surmised what the bird attack rates on the ZIP code and county levels would be for the same 1000 epidemics as follows: as soon as any farm in a ZIP code or county was infected (passed into the Exposed or Latent stage) in a single epidemic, we considered the entire ZIP code or county to be exposed from that time step and to remain exposed for the assumed latent period, then become infectious for the assumed infectious period, then become recovered. These computations generated the second and third rows of Fig. 6 that show the bird attack rates for the ZIP-code and county levels, as indicated on the graphs.

Fig. 7 is analogous to Fig. 6, displaying the unit attack rates of the 1000 epidemics rather than the bird attack rates. By the unit attack rate of an epidemic we mean the fraction of spatial units (i.e., farms, ZIP codes, or counties) that are infected during an epidemic.

All three histograms in Fig. 7 show a bimodal distribution with a clear separation between the peaks on the farm and ZIP-code histograms. On the farm level we labeled an epidemic as mild if it belonged with the left peak in the top histogram of Fig. 7 (529 epidemics) and severe if it belonged with the right peak (471 epidemics). The same criterion on the ZIP-code level resulted in the same epidemics as being either mild or severe. On the county level, however, the bottom histogram of Fig. 7 does not have a clear separation between its two peaks, and so we arbitrarily chose those epidemics to the left of the low point as mild (507 epidemics) and the remainder as severe (493 epidemics).

Fig. 8 displays a bar chart of the average, over the severe epidemics out of 1000 simulations, of the number of newly infectious birds (left) or spatial units (right) in each of the time steps for the three spatial levels. Recall that the time step was one day on the farm level, five days on the ZIP-code level, and ten days on the county level. A ZIP code or county was counted as being newly infected in the time step in which any farm within it was first infected. Notice that the newly infected spatial unit charts on the right always start with one since only the one index farm is newly infected in the first time step.

All three bar charts exhibit the characteristic behavior of such epidemic curves: they build up to a peak rather quickly and then slowly drop down to zero. The peaks, however, are reached sooner as the spatial units increase in size.

Figs. 9-11 display: another important evaluation of an epidemic: the probability that a particular spatial unit will be infected for a given index unit. In each of the three maps of Pennsylvania, each spatial unit
is shaded according to the fraction of times it was one of the infected units among the severe epidemics. On the farm level, sizes of the farm disks are also drawn in one of three sizes according to their probability of being infected.

Notice that many of the ZIP codes (Fig. 10) and counties (Fig. 11) are shaded white, indicating that these spatial units were never infected in any of the severe epidemics among the 1000 simulated. Mathematically speaking, however, every spatial unit has a nonzero probability of being infected in one time step if any other spatial unit is infectious in that time step.

Our estimated epidemic parameters

A good test of how much information is lost by collecting data on larger spatial scales is how well the epidemic parameters estimated from such data model ongoing or future epidemics. We previously considered the question of parameter estimation (Rorres et al., 2010), but only for data collected on the farm level. We considered three different methods for determining the parameters depending on what data were collected. Here we will only apply one of the methods, the most accurate one, which is a maximum likelihood estimation. The method requires that we know which units were infectious in each time step and which susceptible units were and were not infected within each time step. When this amount of data is available, a maximum likelihood estimation (MLE) of the epidemic parameters can be formulated.

To describe this estimation technique, let $A(k)$, as before, be the set of indices of those units that were infectious at the beginning of the $k$th time step, $k = 1, 2, ..., K$, where $K$ is the last time step for which data were collected. Usually, $K$ is the last time step of a completed epidemic, but it could also be the present time step of an ongoing epidemic. Next, let $B(k)$ and $C(k)$ be the sets of indices of the susceptible units that were and were not infected during the $k$th time step, respectively. Then for a given pair of values of $\delta$ and $\rho$ the probability $L(\delta, \rho)$ of the observed pattern of infectious, non-infected, and newly infected units over the $K$ time steps, based on our probabilistic model of the spread of the epidemic, is

$$L(\delta, \rho) = \prod_{k=1}^{K} \left\{ \prod_{j \in A(k)} P_{jk}(\delta, \rho) \prod_{j \in B(k)} \left(1 - P_{jk}(\delta, \rho)\right) \right\}$$  \hspace{1cm} (6)

where $P_{jk}(\delta, \rho)$ is as defined in Eq. (5). Then, by definition, the maximum likelihood estimates for $\delta$ and $\rho$ are those values for which $L(\delta, \rho)$ is maximized.

In Fig. 12 we display these epidemic-parameter estimates for the three spatial scales for some of the control epidemics of the previous section. We used 749 of the control epidemics for the farm-level estimates, 421 for the ZIP-code level estimates, and 451 for the county-level estimates. We discarding those control epidemics at each level that did not converge using a MATLAB™ program for maximizing $L(\delta, \rho)$, usually because the epidemics were too mild to yield acceptable results.

The large circles are the sample means of the three sets of estimates and their numerical values are given in Table 1. We used the arithmetic mean of the $\rho$ estimates and the geometric mean of the $\delta$ estimates.

Because the epidemics from which the MLE estimates were derived were generated from our mathematical model on the farm level using the epidemic parameters $\delta = 5.8 \times 10^{-5}$ and $\rho = 2.50$, it is no surprise that the sample means of the farm-level estimates are reasonably close to the original values. The estimated parameters on the ZIP-code and county levels, however, need not be expected to match the farm-level estimates since different time steps are used for those two levels. What we expect of these estimates is that when inserted into the mathematical model, the resulting simulations will have properties similar to those described in Figs. 6 and 7.

Fig. 13 displays four graphs that illustrate the behavior of our model on the farm level using the two farm-level parameter estimates in Table 1. The two top graphs of the bird attack rates of 1000 simulations are analogous to the two top graphs of Fig. 7. The agreement is excellent, although the jump in Fig. 7 occurs at 0.529 while it is at 0.622 in Fig. 13. Thus the estimated parameters yield a model in which there are fewer severe epidemics than among the control simulations.

The bottom two graphs in Fig. 13 compare very well with their respective analogs of the top-right graph of Fig. 8 (expected daily newly infectious farms of severe epidemics) and Fig. 9 (probability of infection by farm for severe epidemics). All of this is in agreement with our previous conclusions (Rorres et al., 2010) about the accuracy of this method for estimating the two epidemic parameters.

Fig. 14 shows four graphs that describe the typical behavior of epidemics generated by using the ZIP-code level geographical data and the ZIP-code level MLE estimates of $\delta$ and $\rho$ in Table 1. The two top graphs of the bird attack rates of 1000 simulations are analogous to the two middle graphs of Fig. 7. The agreement is again excellent, with a jump in the farm attack rate occurring at 0.529 among the 1000 control epidemics and at 0.457 among the 1000 simulations shown in the top-left graph of Fig. 14.

As with Fig. 13, the bottom two graphs in Fig. 14 compare very well with their respective analogs of the middle-right graph of Fig. 8 (expected daily newly infectious farms of severe epidemics) and Fig. 10 (probability of infection by farm for severe epidemics).

Fig. 15 displays the graphs associated with collected data on the county level and using the analogous county-level mathematical model. Because there are only 67 counties, the results were much cruder than on the farm and ZIP-code levels. We needed to run 100,000 simulated epidemics, rather than 1000, to get a reasonable number of epidemics in which the epidemic spread outside of the index county. Specifically, of 100,000 simulated epidemics 97,661 did not spread beyond the index county, 2330 spread to one other county, 149 spread to two other counties, and 7 spread to three other counties. Of the 67 counties in Pennsylvania, only 7 other

<table>
<thead>
<tr>
<th>Spatial Scale</th>
<th>$\delta$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farm level</td>
<td>$6.9 \times 10^{-5}$</td>
<td>2.55</td>
</tr>
<tr>
<td>ZIP-code level</td>
<td>$7.4 \times 10^{-8}$</td>
<td>1.59</td>
</tr>
<tr>
<td>County level</td>
<td>$4.5 \times 10^{-5}$</td>
<td>2.73</td>
</tr>
</tbody>
</table>

Table 1: Sample means of the maximum likelihood estimates of the system parameters $\delta$ and $\rho$ for the three spatial scales.
than the index county were ever infected among the 100,000 simulations. Consequently, the epidemic parameters determined by the 1000 control epidemics did not generate epidemics consistent with the control epidemics themselves. Thus with an ongoing epidemic for which only county-level data is collected, our model cannot be expected to accurately predict the future course of the epidemic.

Discussion

Our objective was to see how epidemics simulated on the farm, ZIP-code, and county levels behave. We generated a set of 1000 simulated control epidemics on the farm level using a specific set of epidemic parameters and then examined them on the three spatial levels. We then estimated the epidemic parameters on the three
spatial levels and generated more epidemics at the various levels to see how they compare with the control epidemics.

The results, as displayed in Figs. 13–15, suggest that aggregating the farms within ZIP codes produces results as good as those in which the farms are not aggregated. Going from 7043 farms to 805 ZIP codes seems to preserve those features of an epidemic that are useful for the development of vaccination or culling strategies to prevent or control an ongoing or future epidemic.

However, aggregating the farms within 67 county units did not produce satisfactory results. The simulated epidemics that arose when inserting the MLE estimates for the epidemic parameters did not display the characteristics of the farm-level epidemics. Apparently the dynamics of an epidemic within a county are too important to simply describe as one in which all of the farms are simultaneously in a latent stage followed abruptly by an infectious stage. Too much of the spatial detail of our model is lost when only 67 spatial units are considered. It appears that modelers must have data available at least on the ZIP-code level, and ideally on the farm level, to realistically deal with future or ongoing epidemics using modern stochastic, spatial models.

**Acknowledgments**

The project described was supported by award number 5U01GM-076426 from the National Institute of General Medical Sciences (NIGMS). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

**References**


**Fig. 15.** Epidemic graphs using the county-level model and the MLE estimates for the parameters.