Modeling and Real-Time Prediction of Classical Swine Fever Epidemics

Ronald Meester, Jan de Koning, Mart C. M. de Jong, and Odo Diekmann

1Mathematical Institute, University of Utrecht, P.O. Box 80010, 3508 TA Utrecht, The Netherlands
2Institute for Animal Science and Health, Department of Immunology, Pathobiology and Epidemiology, P.O. Box 65, 8200 AB Lelystad, The Netherlands
*email: meester@math.uu.nl

SUMMARY. We propose a new method to analyze outbreak data of an infectious disease such as classical swine fever. The underlying model is a two-type branching process. It is used to deduce information concerning the epidemic from detected cases. In particular, the method leads to prediction of the future course of the epidemic and hence can be used as a basis for control policy decisions. We test the model with data from the large 1997-1998 classical swine fever epidemic in The Netherlands. It turns out that our results are in good agreement with the data.

KEY WORDS: Branching process; Classical swine fever; EM algorithm; Maximum likelihood; Modeling of epidemics; Prediction of epidemics.

1. Introduction and Background

In 1997-1998, The Netherlands suffered from a huge outbreak of classical swine fever (CSF). For a description of this epidemic, see Stegeman et al. (1999a). In order to reduce the transmission of the CSF virus, it turned out to be necessary to implement an increasing number of measures. The decision whether or not measures must be implemented of course highly depends on expectations and predictions concerning the actual epidemic. Given the fact that any decision is based on predictions, the prediction process deserves scrutiny. The only hard data that is available is the number of newly detected cases each week. To calculate a prediction from such data, we shall take into account that only some of the infected herds have been discovered and that the number of infected herds in the past depends on the number of undetected infectious herds that were present then. To extrapolate the stochastic processes of infection and detection from the data, we shall use a mathematical model.

A recent attempt to formulate a model for the evolution of such an epidemic can be found in Stegeman et al. (1999b). In order to set the framework for our own contribution, we start the discussion with a short description of their set-up. The unit of time in their model is one week. The parameters in their model are $N$, the total number of herds; $\beta$, the infection rate; and $\alpha$, the depopulation rate. The variables are $S$, the number of susceptible herds; $C$, the number of newly infected herds per week; $D$ the number of infected herds depopulated per week; and $I$, the number of infected herds. It is assumed that when the virus is discovered in a herd, this herd is depopulated immediately, i.e., the depopulation rate is simply the discovery rate. The relations are the following:

$$C = \frac{\beta IS}{N},$$
$$D = \alpha I.$$

Setting

$$R_0 := \frac{\beta}{\alpha},$$

the goal is to estimate $R_0$ since $R_0 > 1$ corresponds to a potentially explosive situation (see de Jong, 1995). The variable $S$ is assumed to be constant and equal to $N$. Since $\alpha = 1/T$, where $T$ denotes the average period that a herd is infectious, estimation of $R_0$ boils down to estimating the unknowns $T, C,$ and $I.$ (Note that $D$ and $N$ are observable.) In essence, the following method is employed: When a herd is found infected and therefore depopulated, the moment it became infected (which lies in the past, of course!) is drawn from a certain probability distribution, independently from all other herds. In retrospect, i.e., after the epidemic has been going on for a long time, this then gives enough information to compute the number of infected herds at each unit of time, and this in turn yields enough information to estimate all the parameters.

This idea has two serious drawbacks. In the first place, although it can be argued that infectious periods of different herds are indeed independent, when we look backward in time, this independence fails. Indeed, when we know in retrospect that a certain herd was infected at a certain moment in the past, this will have an effect on the probability that other herds were infected around that time. It seems that one does not use all the available information when infection times in...
the past are considered as just shifted i.i.d. random variables. Second, this model yields, in fact, a realization of the course of the epidemic according to the distributions chosen. But to draw such a realization of, say, the first month of the epidemic, we need information about the epidemic long after this first month. This means that we cannot predict in real time.

In general, little has been done on estimation of parameters in an ongoing epidemic using data from that same epidemic. In Becker (1974), some aspects of this problem are discussed under certain assumptions that are inappropriate for our present context, e.g., the assumption of a short infectious period.

We shall model the evolution of the epidemic with a two-type branching process in which dependencies as described above are implicit. The requirement for real-time prediction turns out to lead to an interesting mathematical question. The point is that the data of the epidemic consist only of the depopulated herds. We have no direct information about the number of infected herds, and statements about this number must be inferred from the known number of depopulated herds in the weeks so far. We shall see that such statements can indeed be made and justified in our model.

The organization of the article is as follows. In the next section, we introduce the model and explain the idea behind our approach. In Section 3, we test our method with the data from the 1997–1998 CSF epidemic in The Netherlands. In Section 4, we discuss justification of the model, methodology, results, and applicability. We end with an appendix, in which some mathematical details can be found.

2. The Model and the Mathematical Formulation of the Problem

We model the progress of the epidemic with a discrete-time multitype branching process. General references for such processes are Harris (1963) and Athreya and Ney (1972). As in Stegemann et al. (1999b), the unit of time is one week. In the branching process description, there is an unlimited number of susceptible herds, and we focus entirely on the infected and the newly depopulated herds. We shall adopt the following terminology: In each week \( k \), a herd is said to be of type \( i \) if it has been infected in week \( k \) or before and has not yet been depopulated; a herd in week \( k \) is said to be of type \( d \) if it is depopulated in week \( k \). The process starts with one herd of type \( i \). The transition from week \( k \) to week \( k+1 \) is as follows: A herd of type \( d \) has no offspring at all in week \( k+1 \). For a herd of type \( i \), the rules are the following: With probability \( 1 - \mu \), it has only offspring of type \( i \) and the number of such ‘children’ is equal to one plus a number chosen according to a right-truncated Poisson distribution with parameter \( \lambda \); with probability \( \mu \), it has one child of type \( d \) and a number of children of type \( i \), where this number is chosen according to a right-truncated Poisson distribution with parameter \( \lambda/2 \) (and the same truncation level as above).

In words, the model boils down to the following description: Every week, there is a fixed probability (the parameter \( \mu \), which corresponds to the parameter \( \alpha \) in Section 1) that a herd that has been infected is discovered. This herd is then immediately depopulated. If a herd is not discovered, it infects a random number of other herds, with a truncated Poisson distribution. (The parameter \( \lambda \) of this Poisson distribution corresponds to the parameter \( \beta \) in Section 1.) The extra one in the description is the original infected herd itself. If a herd is discovered, it still may have infected other herds earlier in that same week. The fact that we then use the parameter \( \lambda/2 \) for the Poisson distribution reflects the idea that, on average, detection will take place halfway through the week, in which case the herd has only half as much time to infect other herds than if it were not detected. (The careful reader will recognize the fact that letting the end of the infectious period be uniformly distributed over a week does not lead to an exact Poisson-distributed number of infections when we think of the infection process as a Poisson process in time; we do not think that this is very important.) Note that there are 3 d.f. in this model: the parameters \( \lambda \) and \( \mu \) and the truncation level of the Poisson distributions. We shall choose the truncation level beforehand and estimate \( \lambda \) and \( \mu \) from the data.

Let us turn to the questions of interest. We only observe type \( d \) herds. In particular, we do not know when the epidemic started. Our data, therefore, consists of a sequence of numbers \( (d_1, d_2, \ldots, d_m) \), say, where \( d_i \) is the number of detections in week \( i \). Based on this information, we want to estimate the parameters \( \lambda \) and \( \mu \) and to obtain an indication of the precision of these estimates. We also want to make probabilistic predictions of the future of the epidemic based on these estimates. In the rest of this section, we describe the method.

We call the first week in which we observe depopulation week 1. (The week before this is then called week 0.) The (unknown) number of type \( i \) herds in week \( n \) is denoted by \( X_n \) and the number of type \( d \) herds by \( Z_n \). The data we have tells us that \( Z_1 = d_1 \) for \( i = 1, \ldots, m \). The event \( \{Z_n = d_n, Z_{n+1} = d_{n+1}, \ldots, Z_m = d_m\} \) is denoted by \( D_n^m \) for \( n = 1, \ldots, m \). We write \( D = D_1^m \).

Estimating the parameters. For a certain choice of the parameters, we want to find the probability of the observed data. Once we know how to do this, we can use the result to estimate \( \lambda \) and \( \mu \) via a maximum likelihood procedure. Fix some \( \lambda \) and \( \mu \). We distinguish the following steps:

Step 1. We recursively find \( P(D_n^m \mid X_{n-1} = i) \) for \( n = m, m-1, \ldots, 1 \) (in that order) and for all \( i \). With \( n = 1 \), this means that we know \( P(D \mid X_0 = i) \).

Step 2. We find the exact distribution of \( X_0 \).

Step 3. With steps 1 and 2, we can compute \( P(D) = \Sigma_i P(D \mid X_0 = i)P(X_0 = i) \), which can be used to estimate \( \lambda \) and \( \mu \) as explained above.

Prediction of future development. We distinguish between two approaches for future prediction.

Using the previous steps, we now know \( P(X_0 = i \mid D) \) since

\[
P(X_0 = i \mid D) = \frac{P(X_0 = i)P(D \mid X_0 = i)}{P(D)}
\]

We next compute recursively \( P(X_n = i \mid D) \) for \( n = 1, 2, \ldots, m \) (in that order) and for all \( i \). In the Appendix, we show how to perform these recursive computations. This last step yields, in particular, the correct conditional distribution (i.e., conditioned on \( D \), which is all the information we have) of the number of type \( i \) herds in week \( m \). Since our model is a Markov chain, this conditional distribution is enough to compute the distribution of every possible statistic about the future of the process. Note that all the above conditional distributions really depend on the whole history of the process: We have used...
all information available to us. Also note that we can predict in real time: In week \( m \), we can make predictions for the next week, \( m+1 \).

Another approach is to simply compute \( R_0 \), i.e., the expected number of herds that are infected by one particular herd. With probability \( \mu \), the expected number of secondary cases equals \( \lambda/2 \), and with probability \( 1-\mu \), it will be \( \lambda \) (for the first week) plus \( R_0 \) (for all subsequent weeks). It follows that

\[
R_0 = \frac{\lambda\mu}{2} + (1-\mu)(\lambda + R_0),
\]

and we obtain

\[
R_0 = \frac{\lambda(2-\mu)}{2\mu}.
\]

The epidemic will almost surely die out when \( R_0 \leq 1 \) and possibly survive otherwise.

Uncertainty. There are a number of ways to obtain an indication about the uncertainty of the estimates. The most obvious way to do this is to use a bootstrap procedure, like Efron’s quantile method (cf., Van der Vaart, 1998). For us, there is a practical problem with this method though. Computing an ML estimate in our setting requires a lot of computing time (in the order of 5-8 hours on a Sun Sparc station). This means that a reasonable number of ML estimates cannot be obtained in a day, say. Since we insist on real-time prediction, this is unacceptable.

We resolve this problem as follows: Instead of computing the ML estimate, we use a modified EM algorithm (cf., McLachlan and Krishnan, 1997) to obtain an approximation of the ML estimates of the parameters. Note that, under certain circumstances, the EM algorithm is known to converge to the exact ML estimate. As we will see later, for computational purposes, we modify the algorithm. This means that our results will not be exact, even if the EM algorithm would give exact results. The starting value of the EM algorithm will be the previously computed ML estimate based on the data. The procedure runs as follows.

We start by generating on a computer many (100, in our case) realizations of the epidemic using the previously determined ML estimations based on the data. Consider the first such generated realization. Using the full information of the generated data, it is very easy to estimate (with a maximum likelihood procedure) the parameters \( \lambda \) and \( \mu \). Call these estimates \( \lambda_0 \) and \( \mu_0 \). In our case, we cannot quite compute the appropriate conditional expectation, i.e., conditional on the series of detected herds. Instead, we approximate this conditional expectation through simulation as follows:

Step 1. Using \( \lambda_0 \) and \( \mu_0 \), we generate many (50, in our case) realizations of the process conditioned on the series of detected herds in the original (first) simulation (in the Appendix, we shall explain how to do this) as follows:

Step 2. In each of these 50 realizations, we again have full information, and we estimate the parameters with an ML procedure.

Step 3. We then define \( \lambda_1 \) and \( \mu_1 \) as the average values of these 50 estimates.

Note that, in step 3, we cheat a little bit. We compute the average of the maximizing parameter values, where we should in fact take the parameter value that maximizes the average log likelihood; we replace the maximizer of the average by the average of the maximizers. This is done for computational reasons—the latter is much easier to compute.

Steps 1–3 are repeated (each time conditioned on the same number of detected herds each week) with \( \lambda_1 \) and \( \mu_1 \), etc., until we observe convergence of the estimates \( \lambda_k \) and \( \mu_k \). The convergence criterion we used is that we stop as soon as the first two decimals remain unchanged under an iteration. This value is then our final estimate. In this way, we obtain 100 estimates, and a confidence region for the parameters can now be obtained by disregarding the 10% extreme estimates, as usual. With this method, we can obtain an approximate 90% confidence interval in about 12 hours. The accuracy of the approximation and performance with actual data of the EM algorithm will be addressed in the next sections.

3. Results

We have tested the method with the data from the 1997–1998 epidemic in The Netherlands. These data are available at the website www.minlnv.nl/varkenspest/. With our unit of time, the data is given in Table 1.

In Stegeman et al. (1999b), five periods characterized by different control measures were described. A change in a control measure results in a potential change of the parameters. We therefore have 2 parameters per period, a total of 10 parameters. In our labeling of time, the first period consists of week 1 only. The second period is weeks 2–11, the third weeks 12–19, and the fourth period is weeks 20–28. The fifth period starts at that point and lasts until the end.

When computing estimates, we have to take a maximum for the number of undetected infected herds. For this maximum number, we chose 149. This number turned out to be convenient and big enough compared with the numbers involved. When estimating the parameters, we cannot compute \( P(D) \) for all feasible values of the parameters. Instead, we put a grid with mesh-width 0.05 or 0.025 on the multidimensional parameter space, and maximize \( P(D) \) on this grid via a standard search procedure. We now describe our results.

3.1 Estimates and Predictions

Consider the data up to and including the first time we had four zeroes in a row, i.e., until week 47. In Table 2, we give the

<table>
<thead>
<tr>
<th>Table 1</th>
<th>The data of the 1997–1998 CSF epidemic in The Netherlands</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>1</td>
</tr>
<tr>
<td>( Z_n )</td>
<td>10</td>
</tr>
<tr>
<td>( Z_n )</td>
<td>11</td>
</tr>
<tr>
<td>( Z_n )</td>
<td>19</td>
</tr>
<tr>
<td>( Z_n )</td>
<td>21</td>
</tr>
<tr>
<td>( Z_n )</td>
<td>31</td>
</tr>
<tr>
<td>( Z_n )</td>
<td>8</td>
</tr>
<tr>
<td>( Z_n )</td>
<td>41</td>
</tr>
<tr>
<td>( Z_n )</td>
<td>1</td>
</tr>
<tr>
<td>( Z_n )</td>
<td>51</td>
</tr>
</tbody>
</table>
approximated 90% confidence intervals and the ML estimates for the parameters. The corresponding confidence intervals for $R_0$ are given in Table 3. One remark about these last intervals is important. We deduced the confidence intervals for $R_0$ from the corresponding intervals from $\lambda$ and $\mu$. If one is primarily interested in $R_0$, one could (and should maybe) deduce the confidence interval for $R_0$ directly by viewing $R_0$ as a parameter of the model.

We also computed (with the ML estimates) the estimated probability that $X_{47} = 0$, i.e., the probability that the epidemic is over; we found $P(X_{47} = 0) = 0.8$. Hence, there is an estimated probability of 0.2 that the epidemic is not yet over.

Next, consider the data of the first 6 weeks only. The corresponding estimates and confidence intervals can be found in Table 4. We also computed the distribution of $X_6$ and $Z_7$ given $Z_1 = d_1, \ldots, Z_6 = d_6$ with the ML estimates in Table 4. The distribution of $X_6$ turns out to be very much spread out: $P(X_6 = i) > 0.01\) for all $14 \leq i \leq 41$.

The conditional distribution of $Z_7$ given $Z_1 = d_1, \ldots, Z_6 = d_6$ can also be computed. This distribution is somewhat less spread out. For the ML estimates of the parameters above, we find that $P(Z_7 = i) > 0.1$ for all $3 \leq i \leq 8$. (In reality, $Z_7$ was equal to nine.)

3.2 The EM Algorithm Applied to the Actual Data to Compare Its Performance to True ML Estimation

As anticipated in the previous section, we have tested the algorithm on the actual data. One point of concern here is the choice of the starting values. We have performed the algorithm for three different starting values, two fairly close to the exact ML estimates and one quite far away. We report the results for $\lambda$ (see Table 5). In Table 5, the last three rows correspond to starting values that are, respectively, 1.25, 0.75, and 4 times the exact ML values.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>ML estimates and approximated 90% confidence intervals based on weeks 1-47</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>2</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>[0.56, 0.63]</td>
</tr>
<tr>
<td>$\mu$</td>
<td>[0.39, 0.43]</td>
</tr>
<tr>
<td>$R_0$</td>
<td>[0.40, 0.49]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Approximated 90% intervals for $R_0$ based on weeks 1-47</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>2</td>
</tr>
<tr>
<td>$R_0$</td>
<td>[1.02, 1.30]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4</th>
<th>ML estimates and 90% confidence intervals for the parameters and $R_0$ based on the first 6 weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>1</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>[2.2, 3.4]</td>
</tr>
<tr>
<td>$\mu$</td>
<td>[0.40, 0.49]</td>
</tr>
<tr>
<td>$R_0$</td>
<td>[3.4, 6.8]</td>
</tr>
</tbody>
</table>

Discussion

4.1 Model Justification

We neglect the geographical locations of the herds. In principle, we could incorporate geographical structure by increasing the number of types. However, such a procedure would increase the complexity of the model, and it is at least questionable whether it would lead to a significant improvement. But we stress that our method is robust and could in principle be applied to a more complex model, incorporating this extra ingredient.

The fact that we truncate the Poisson distributions is of little importance for the outcome. This was checked by repeating the procedure for different values of the truncation level. We chose five as the default truncation level; higher truncation levels hardly made any difference, and this level guarantees that the computations are manageable, i.e., do not take too much time. The choice of the Poisson distribution itself is arbitrary and natural at the same time. If there are good reasons, we could choose another distribution. Our method will work with any distribution with finite support.

One could question the idea behind the use of a branching process. If the total number of infected herds is of the same order as the total population of herds, then clearly a branching process cannot be a reasonable description. In the case of the huge epidemic in The Netherlands in 1997–1998, the total number of infected herds was 429, the total population of herds 21,500. These numbers seem to suggest that a branching approach in fact be meaningful. Yet one should realize the deliberate simplification discussed above: If contacts are primarily local, one still might have to take local exhaustion of susceptible herds into account.

The choice of a week as the unit of time is motivated by the special characteristics of CSF. Within a week after infection, most herds will become contagious, thereby increasing the complexity of the model. A longer unit of time would introduce the possibility that herds are infected and depopulated at the same time, and this is certainly undesirable. A week seems a reasonable compromise.

The parameters of the model will change when measures are implemented. For instance, after discovery of the first case of CSF, it is reasonable to expect that the discovery rate $\mu$ goes up. In addition, measures are taken that should

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Results of the modified EM algorithm on the actual data with different starting points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>2</td>
</tr>
<tr>
<td>Exact ML result</td>
<td>0.6</td>
</tr>
<tr>
<td>EM results (factor 1.25)</td>
<td>0.7</td>
</tr>
<tr>
<td>EM results (factor 0.75)</td>
<td>0.6</td>
</tr>
<tr>
<td>EM results (factor 4.00)</td>
<td>—</td>
</tr>
</tbody>
</table>
decrease the infection rate $\lambda$. Therefore, when applying this model, one should take into account different time periods characterized by differences in control measures and estimate the parameters in each such period. Obviously, there are counteracting forces at work here, as precision of estimates benefits from longer time series.

4.2 Comments to Modified EM Algorithm

The number of simulations we use is such that we can cope with the computational burden in real time. At the same time, the confidence intervals are not inappropriately large.

When we applied the modified EM algorithm, convergence (in the sense described above) typically took place around the 10th iteration.

One could ask whether or not estimation by the modified EM method is a convenient alternative to exact ML estimates. There are two reasons why we do want to use the exact ML method. In the first place, the ideas behind our exact computation are interesting in their own right. It is not a priori clear that such an exact calculation is practically possible at all, and the method is interesting from a theoretical point of view. Second, the ML estimates are a very natural starting point for the EM algorithm. It is not clear how one should choose an initial value without reference to the exact ML estimates, and in the previous section, we saw that bad starting points can lead to convergence problems.

4.3 Results

We discuss a number of issues related to the results of Section 3.

After the 4 weeks 44–47 without detection, the probability that the epidemic is not yet over is 0.2. This is quite a high number at first sight, and in particular it might be too big to cancel control measures. Reality proved this to be wise.

Some estimates for the parameters may seem counterintuitive, e.g., the fact that the estimates for $\mu$ in the fourth and fifth period are lower than in the second and third. Interestingly, Stegeman et al. (1999b) also observed this in their estimates and gave an explanation, which we do not repeat here. On the other hand, the endpoints of the confidence intervals for $R_0$ are in fact decreasing in time, as they should be. Apparently, the decrease in $\lambda$ compensates for the decrease in $\mu$. Note that only in the last period is $R_0$ smaller than one throughout the whole confidence interval.

The distribution of $X_6$ is very much spread out. This suggests that the structure of the problem is such that, even with fixed (and known) parameters, the uncertainty in the number of infected herds around is very large. A large degree of uncertainty seems to be an intrinsic aspect of the model. This may be important information for decision makers.

4.4 Applicability

It seems that the model describes the epidemic in a satisfactory way and that reasonable probabilistic predictions of future development can be obtained. One question is of course how to use this algorithm in practice. Decisions whether or not to implement control measures depend on many things, e.g., the risk one is willing to take. At some point, our algorithm might give probability 0.9 that the epidemic is over in 2 weeks. Is this high enough? This is not a mathematical question. But estimates for various statistics of future happenings can lead to a fairly complete picture of what can be expected, and on these, one can base political choices. In this spirit, we are currently making our method operational.

RÉSUMÉ

Nous proposons une nouvelle méthode d'analyse du déclenchement d'une maladie infectieuse telle que la fièvre porcine. Le modèle sous-jacent est un processus de branchement à 2 espèces. Il est utilisé pour obtenir une information sur l'épidémie à partir des cas détectés. En particulier, la méthode permet de prédire le développement de l'épidémie, et peut ainsi être servir de base aux décisions politiques de contrôle. Nous testons le modèle avec des données provenant de la grande épidémie de fièvre porcine classique survenue en 1997–1998 aux Pays-Bas. Nous observons que nos résultats sont en bon accord avec les données.

REFERENCES


Received April 2000. Revised July 2001. Accepted August 2001.

APPENDIX

Estimating the Parameters

Step 1. We assume that we have values for $\lambda$ and $\mu$, but we will not express this in our notation. We define, for $n = 0, \ldots, m - 1$ and all $i$,

$$P_{n,j} := P(D_{n+1}^+ \mid X_n = j)$$

and

$$f(i, d) = \binom{i}{d} \mu^d (1 - \mu)^{i-d}.$$

Note that $f(i, d) = P(Z_n = d \mid X_{n-1} = i)$. Furthermore, we
define
\[ c(i, j, d) := P(X_n = j | \{X_{n-1} = i\} \cap \{Z_n = d\}). \]

Note that this last probability does indeed not depend on \( n \).

Also note that \( c(i, j, d) \) can be computed in an elementary way. We start by noting that \( P_{m-1,i} = f(i, d_m) \).

To compute \( P_{n,i} \) for values of \( n < m - 1 \), we use backward recursion in \( n \). To this end, observe that
\[ P_{n,i} = P \left( \{Z_{n+1} = d_{n+1}\} \cap \mathcal{D}_{n+2}^+ | X_n = i \right) \]
\[ = \left( P \left( \mathcal{D}_{n+2}^+ | \{Z_{n+1} = d_{n+1}\} \right) \cap \{X_n = i\} \right) \times P \left( Z_{n+1} = d_{n+1} | X_n = i \right). \]

The second term in the last expression is equal to \( f(i, d_{n+1}) \).
The first term is equal to \( \sum_j c(i, j, d_{n+1}) P_{n+1,j} \), leading to
\[ P_{n,i} = f(i, d_{n+1}) \sum_j c(i, j, d_{n+1}) P_{n+1,j}. \]

This enables us to compute in a (backwards) recursive fashion all numbers \( P_{n,i}, n = m - 1, \ldots, 0 \) and all \( i \).

**Step 2.** To compute the (unconditional) distribution of \( X_0 \), we define, for \( i \geq 1 \), \( Q_i \) as the probability that, starting with \( i \) infected herds, the first detection takes place no later than the first infection. Furthermore, \( r_k,n \) is the probability that, starting with \( k \) infected herds, we eventually arrive at exactly \( n \) infected herds without detecting any herds up to and including that point. Note that
\[ P(X_0 = i) = r_{1,i} \cdot Q_i. \]

Theorem 1 tells us how to compute \( Q_i \). Theorem 2, together with the fact that \( r_{n,n} = 1 \), tells us how to compute \( r_{1,i} \). Hence, we can compute the exact distribution of \( X_0 \).

**Theorem 1:** For all \( i \), we have
\[ Q_i = \frac{1 - f(i, 0)}{1 - c(i, i, 0) \cdot f(i, 0)}. \]

**Proof.** Suppose that there are \( i \) infected herds in a particular week. We look at the events that could happen next week. The probability that we detect at least one herd is given by
\[ P(Z_{n+1} > 0 | X_n = i) \]
\[ = 1 - P(Z_{n+1} = 0 | X_n = i) = 1 - f(i, 0). \]

The probability that there is neither detection nor infection next week is given by
\[ P(X_{n+1} = i | \{Z_{n+1} = 0\} \cap \{X_n = i\}) \]
\[ \times P(Z_{n+1} = 0 | X_n = i) = c(i, i, 0) \cdot f(i, 0). \]

When the latter event happens, the conditional probability of getting detection no later than infection has not changed. This leads to \( Q_i = 1 - f(i, 0) + c(i, i, 0) \cdot f(i, 0) \cdot Q_i \).

**Theorem 2:** For all \( n \) and for all \( k < n \), we have
\[ r_{k,n} = f(k, 0) \sum_{j=k}^{n} c(k, j, 0) \cdot r_{j,n}. \]

**Proof.** The probability \( r_{k,n} \) is the probability that we get exactly \( n \) undetected infected herds without any detection given we have \( k \) such herds now. To compute \( r_{k,n} \), we look at the possible situations next week. We have \( k < n \), so if we detect one or more herds, we will not get \( n \) undetected herds without any detections simply because we do not have \( n \) undetected herds and we do detect a herd.

The other possibility is that we do not detect a herd. This happens with probability \( f(k, 0) \). If we do not detect any herd, we can get any number \( j \geq k \) undetected herds next week. This happens with probability \( c(k, j, 0) \). If we have \( j \) undetected herds, next week we have probability \( r_{j,n} \) that we will get \( n \) undetected herds without detecting any herds before. This means that
\[ r_{k,n} = f(k, 0) \sum_{j=k}^{n} c(k, j, 0) \cdot r_{j,n}. \]

It is clear that \( r_{j,n} = 0 \) for \( j > n \), and we conclude that
\[ r_{k,n} = f(k, 0) \sum_{j=k}^{n} c(k, j, 0) \cdot r_{j,n}. \]

**Prediction of Future Development**

Given \( P(X_0 = i | D) \) for all \( i \), we next explain how to find \( P(X_{n-1} = i | D) \) for \( n = 1, 2, \ldots, m \) in a recursive fashion, in that order. (Note that the previous recursion went in the other direction.)

In an inductive fashion, suppose we have computed \( P(X_{n-1} = i | D) \) for all \( i \). We use the identity
\[ P(X_n = i | D) = \sum_j P(X_n = i | \{X_{n-1} = j\} \cap D) \cdot P(X_{n-1} = j | D). \]

With Theorem 3, this leads to
\[ P(X_n = i | D) = \sum_j \frac{c(j, i, d_n) \cdot f(j, d_n) \cdot P_{n,i}}{P_{n-1,j}} \cdot P(X_{n-1} = j | D). \]

which is the desired induction step.

**Theorem 3:** For all \( i \) and for all \( 0 < n \leq m \), we have (with \( P_{m,i} = 1 \)),
\[ P(X_n = i | \{X_{n-1} = j\} \cap D) = \frac{c(j, i, d_n) \cdot f(j, d_n) \cdot P_{n,i}}{P_{n-1,j}}. \]

**Proof.** It is easiest to work from right to left. We assume first that \( n < m \). Some algebra shows that
\[ \frac{c(j, i, d_n) \cdot f(j, d_n) \cdot P_{n,i}}{P_{n-1,j}} = P(\{X_{n-1} = i\} \cap \{Z_n = d_n\} | X_{n-1} = j) \]
\[ \times \frac{P(D_{n+1}^{+} | X_{n-1} = j)}{P(D_{n+1}^{+} | X_{n-1} = i)} \times \frac{P(D_{n+1}^{+} | X_{n-1} = j)}{P(D_{n+1}^{+} | X_{n-1} = i)}. \]

When we know \( X_n \), extra information about \( X_{n-1} \), and \( Z_n \) is irrelevant for the conditional distribution of the future. Hence,
\[ P(D_{n+1}^{+} | X_n = i) = P(D_{n+1}^{+} | \{X_n = i\} \cap \{X_{n-1} = j\} \cap \{Z_n = d_n\}). \]
and when we substitute this into the last expression, this expression reduces to

$$
P(D_{n+1}^+ \cap \{X_n = i\} \cap \{X_{n-1} = j\} \cap \{Z_n = d_n\})
P(D_n^+ \cap \{X_{n-1} = j\})
$$

It follows that

$$
P \left( D_n^+ \cap \{X_n = i\} \cap \{X_{n-1} = j\} \right)
$$

Finally, we note that, for the same reason as above,

$$
P(X_n = i \mid \{X_{n-1} = j\} \cap D)
$$

and we are done in the case \( n < m \). When \( n = m \), the proof is simpler and is left to the reader.

**Uncertainty**

We explain how to draw a realization of the epidemic conditioned on \( D \). We know from the previous results how to compute \( P(X_0 = i \mid D) \) for all \( i \). This allows us to choose a value \( x_0 \) for \( X_0 \) according to this conditional distribution, i.e., according to the law of \( X_0 \) given \( D \). In an inductive fashion, once we have selected \( x_0, x_1, \ldots, x_{n-1} \), we can use Theorem 3 to compute \( P(X_n = i \mid \{X_{n-1} = x_{n-1}\} \cap D) \) for all \( i \) and in particular choose a value \( x_n \) for \( X_n \) according to the correct conditional distribution given \( \{X_{n-1} = x_{n-1}\} \cap D \). According to Theorem 4, we then end up with a random vector \( (x_0, \ldots, x_m) \), which is a realization drawn from the law of \( (X_0, \ldots, X_m \mid D) \). This means we have selected a realization of the epidemic according to the correct conditional distribution, i.e., given \( D \).

**Theorem 4.**

\[
P(X_0 = x_0, \ldots, X_m = x_m \mid D)
= P(X_0 = x_0 \mid D) 
\times \prod_{n=1}^{m} P(X_n = x_n \mid \{X_{n-1} = x_{n-1}\} \cap D).
\]

**Proof.** We know that \( P(X_n = x_n \mid \{X_{n-1} = x_{n-1}\} \cap D) \) is equal to

\[
P(X_n = x_n \mid \{X_{n-1} = x_{n-1}\}, \ldots, \{X_0 = x_0\} \cap D)
\]

because the values of \( X_k \) for \( k < n - 1 \) have no further effect on this probability once we know that \( X_{n-1} = x_{n-1} \). Substituting this in the right-hand side of the statement of the theorem immediately leads to the desired result.