A Method of Determining Where to Target Surveillance Efforts in Heterogeneous Epidemiological Systems:

Supporting Information on the Statistical Model

Alexander J. Mastin\(^1\), Frank van den Bosch\(^2\), Timothy R. Gottwald\(^3\), Vasthi Alonso Chavez\(^2\), and Stephen R. Parnell\(^1\)

\(^1\)Ecosystems and Environment Research Centre, School of Environment and Life Sciences, University of Salford, Greater Manchester, M5 4WT, UK
\(^2\)Computational and Systems Biology, Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ, UK
\(^3\)USDA Agricultural Research Service, 2001 South Rock Road, Fort Pierce, Florida 34945, USA

**Sampling model**

We created a statistically informed sampling model based upon the prevalence of infection (proportion of infected individuals) in each group, a group-specific sample size (\(N_h\) from hosts and \(N_v\) from vectors), and a sampling interval of \(\Delta\) days. As described in the main text, if we assume that first detection occurs at a specified time (\(t_1\)) and that the prevalence at this time (and all previous sampling points) is known, we can estimate to probability of first detection as the product of the probabilities of at least one detection at the sampling point in question, and the probability of no detection in all previous samplings:

\[
P(t_1 | t_0) = \left[1 - \left(1 - \left[\frac{I_h(t_1)}{\rho_h}\right]\right)^{N_h} \cdot \left(1 - \left[\frac{I_v(t_1)}{\rho_v}\right]\right)^{N_v}\right] \prod_{k=1}^{K} \left(1 - \left[\frac{I_h(t_1 - k\Delta)}{\rho_h}\right]\right)^{N_h} \cdot \left(1 - \left[\frac{I_v(t_1 - k\Delta)}{\rho_v}\right]\right)^{N_v}
\]

(1)

In order to simplify this equation further, we make three assumptions:

- The prevalences of infection are low (meaning that we can model this as a Poisson rather than a binomial process).
Sampling is a continuous process rather than occurring at discrete intervals, with a sampling rate of $\frac{N}{\Delta}$ - represented below as $\theta$. This assumption requires the sampling interval to not be too large [1].

The growth in the prevalence in each group is exponential (as would be expected in the early stages of an epidemic).

If we assume that the prevalences of infection prior to first detection $I_h(t_1)$ and $I_v(t_1)$ are low (as would be the case in the early stages of an epidemic), we can model the probability detection at time $t_1$ as a Poisson (rather than a binomial) process by linearising equation 1 using the first terms of a Taylor series expansion:

$$P(t_1 \mid t_0) \approx 1 - \exp \left( - \left( N_h \frac{I_h(t_1)}{\rho_h} + N_v \frac{I_v(t_1)}{\rho_v} \right) \right) \prod_{k=1}^{K} \exp \left( - \left( N_h \frac{I_h(t_1 - k\Delta)}{\rho_h} + N_v \frac{I_v(t_1 - k\Delta)}{\rho_v} \right) \right)$$ (2)

If we assume that sampling occurs a continuous process rather than occurring at discrete intervals, we can simplify equation 2 further to obtain the following equation:

$$P(t_1 \mid t_0) \approx 1 - \exp \left( - \left( N_h \frac{I_h(t_1)}{\rho_h} + N_v \frac{I_v(t_1)}{\rho_v} \right) \right) \cdot \exp \left( \int_{t_0}^{t_1} \left( \frac{1}{\Delta} \left( N_h \frac{I_h(t)}{\rho_h} + N_v \frac{I_v(t)}{\rho_v} \right) \right) dt \right)$$ (3)

And finally, if we assume that the prevalence in each group grows exponentially (and assuming that the prevalence is low), then equation 3 can be further simplified to give:

$$P(t_1 \mid t_0) \approx \left( N_h \frac{I_h(t_0)}{\rho_h} \right) e^{rt_1} + \left( N_v \frac{I_v(t_0)}{\rho_v} \right) e^{rt_1} \cdot \exp \left( - \left( \frac{1}{\Delta} \right) \left( N_h \frac{I_h(t_1-t_0)}{\rho_h} + N_v \frac{I_v(t_1-t_0)}{\rho_v} \right) \right)$$ (4)

Equation 4 describes the probability of first detection at time $t_1$ given pathogen entry at time $t_0$. Since the time of pathogen entry will generally not be known, we can convert this in order to describe the probability of first entry at time $t_0$ given the pathogen was detected at time $t_1$ using Bayes’ theorem [1, 2], to convert this to the probability of first entry at time $t_0$ given the pathogen was detected at time $t_1$. Indicating the ‘sampling rates’ of $\frac{N_h}{\Delta}$ and $\frac{N_v}{\Delta}$, as $\theta_h$ and $\theta_v$, the following equation is obtained:
to convert groups, this stage requires consideration of each group independently. In order
to convert this estimate of the timing of first detection to an estimate of the prevalence at first detection. As this prevalence will vary between the group of interest ($\kappa$), given detection at time $t_1$ ($P(q_0^* | t_1)$), we apply the following transformation to each group:

$$P(q_0^* | t_1) = \left( \frac{1}{r} \right) \left( \frac{1}{q_0^*} \right) \left( \frac{1}{1 - q_0^*} \right) P(t_0 | t_1)$$

(7)

Again, if we assume that the prevalence grows exponentially ($q^* \approx q_0 e^{rt}$), and that the prevalence is low (meaning that $1 - q_0^* \approx 1$), since $\left( \frac{1}{1 - q_0^*} \right) c e^{rt_1} = \left( \frac{I_\kappa(t_0)}{\rho \kappa} \right) e^{rt_1} = q\kappa(t_1)$ the following two formulae are obtained:

$$P(q_h^* | t_1) = \left( \frac{1}{r} \right) \left( \frac{1}{q_h^*} \right) \left( \frac{1}{\nu_h(t_h)} \right) P(t_0 | t_1)$$

(8)

$$P(q_v^* | t_1) = \left( \frac{1}{r} \right) \left( \frac{1}{q_v^*} \right) \left( \frac{1}{\nu_v(t_v)} \right) P(t_0 | t_1)$$

(9)
As described in the main text, equations 8 and 9 represent exponential distributions:

\[ P(q^* | t_1) \sim \text{Exponential} \left( \left( \frac{1}{r} \right) \left( \theta_h + \theta_v \left[ \frac{\nu_h}{\rho_h} \right] \right) \right) \]  
(10)

\[ P(q_v^* | t_1) \sim \text{Exponential} \left( \left( \frac{1}{r} \right) \left( \theta_h \left[ \frac{\nu_v}{\rho_v} \right] + \theta_v \right) \right) \]  
(11)

The mean prevalences at first detection \( E(q^*_h) \) and \( E(q^*_v) \) can be estimated as the inverse of the exponential rate parameter, according to the characteristics of the exponential distributions in Equations 10 and 11:

\[ E(q^*_h) = \frac{r}{\theta_h + \theta_v \left[ \frac{\nu_h}{\rho_h} \right]} \]  
(12)

\[ E(q^*_v) = \frac{r}{\theta_h \left[ \frac{\nu_v}{\rho_v} \right] + \theta_v} \]  
(13)

We can rearrange these to give an estimate of the sampling rate required for detection at a specified mean prevalence in either hosts \( E(q^*_h) \) or vectors \( E(q^*_v) \). Since we are also interested in both the sampling rate from hosts and vectors, this gives us a total of four equations:

The sampling rate required from hosts for detection at a specified mean host prevalence:

\[ \theta_h = \left( \frac{r}{E(q^*_h)} \right) - \left[ \frac{\nu_h}{\rho_h} \right] \theta_v \]  
(14)

The sampling rate required from vectors for detection at a specified mean host prevalence:

\[ \theta_v = \left( \frac{r}{E(q^*_h)} \right) \left[ \frac{\nu_h}{\rho_h} \right] - \left[ \frac{\nu_v}{\rho_v} \right] \theta_h \]  
(15)

The sampling rate required from hosts for detection at a specified mean vector prevalence:

\[ \theta_h = \left( \frac{r}{E(q^*_v)} \right) \left[ \frac{\nu_v}{\rho_v} \right] - \left[ \frac{\nu_h}{\rho_h} \right] \theta_v \]  
(16)

The sampling rate required from vectors for detection at a specified mean vector prevalence:

\[ \theta_v = \left( \frac{r}{E(q^*_v)} \right) - \left[ \frac{\nu_h}{\rho_h} \right] \theta_h \]  
(17)
These are linear equations (containing an ‘intercept’ and a ‘slope’ term), indicating that when sampling to detect a specified mean prevalence (in either hosts or vectors), either hosts or vectors (but not both) should generally be sampled. If this is the case, the ‘intercept’ term alone in equations 14 - 17 will describe the required sampling rate (which we describe as $\Theta_h$ and $\Theta_v$, to distinguish the single group sampling rate from the combined sampling rate, $\theta$, in equations 14-17). We can reformulate these new equations in terms of the expected prevalence at first detection. Using equations 14 and 15, we can describe the mean host prevalence at first detection as a function of the host or the vector sampling rate:

$$E(q^*_h) = \left( \frac{r}{\Theta_h} \right) = \left( \frac{r}{\Theta_v} \right) \left[ \frac{\nu_h}{\rho_h} \right] \left[ \frac{\nu_v}{\rho_v} \right]$$

(18)

And using equations 16 and 17, we can describe the mean vector prevalence at first detection as a function of the host or the vector sampling rate:

$$E(q^*_v) = \left( \frac{r}{\Theta_h} \right) \left[ \frac{\nu_v}{\rho_v} \right] = \left( \frac{r}{\Theta_v} \right)$$

(19)

Both of these formulations simplified down into the following equation, which describes the relative sampling rate required when vectors are exclusively sampled and when hosts are exclusively sampled, for any specified mean prevalence at first detection in either hosts or vectors:

$$\left( \frac{\Theta_v}{\Theta_h} \right) = \left[ \frac{\nu_v}{\rho_v} \right] \left[ \frac{\nu_h}{\rho_h} \right]$$

(20)

The right side of equation 20 can be interpreted as a ‘threshold quantity’. A value of greater than one indicates that the sampling rate required when sampling only from vectors is greater than that required when sampling only from hosts; and a value of less than one indicates the converse. We describe in the main paper how costs can be incorporated into this framework.

References
