Risk Analysis DOI: 10.1111/risa.13592

Estimating Consignment-Level Infestation Rates from the Proportion of Consignment that Failed Border Inspection: Possibilities and Limitations in the Presence of Overdispersed Data

Raphaël Trouvé [1,*] **and Andrew P. Robinson**[2]

ABSTRACT: Introduction of pests and diseases through trade is one of the main socioecological challenges worldwide. Targeted sampling at border security can efficiently provide information about biosecurity threats and also reduce pest entry risk. Prioritizing sampling effort requires knowing which pathways are most infested. However, border security inspection data are often right-censored, as inspection agencies often only report that a consignment has failed inspection (i.e., there was at least one unit infested), not how many infested units were found. A method has been proposed to estimate the mean infestation rate of a pathway from such right-censored data (Chen et al.). Using simulations and case study data from imported germplasm consignments inspected at the border, we show that the proposed method results in negatively biased estimates of the pathway infestation rate when the inspection data exhibit overdispersion (i.e., varying infestation rates among different consignments of the same pathway). The case study data also show that overdispersion is prevalent in real data sets. We demonstrate that the method proposed by Chen et al. recovers the median infestation rate of the pathway, rather than its mean. Therefore, it underpredicts the infestation rate when the data are overdispersed (in right-skewed distributions, the mean is above the median). To allow better monitoring and optimizing sampling effort at the border, we recommend that border protection agencies report all the data (the number of infested units found together with the sample size of the inspection) instead of only that the consignment failed inspection.

KEY WORDS: Biosecurity inspection; live plant imports; overdispersion; quarantine pest; right-censoring

1. INTRODUCTION

The invasion of alien pests and diseases is one of the most important socioecological challenges worldwide. The cost associated with biological invasions are substantial and in the range of 2020 USD \$3–200 billions for different countries,[1] e.g., 2020 USD\$2.6 billion for New Zealand (Giera & Bell,

[1]CEBRA & School of Ecosystem and Forest Sciences, The University of Melbourne, Parkville, Victoria, Australia.
[2]CEBRA & School of Mathematics and Statistics, The University of Melbourne, Parkville, Victoria, Australia.
*Address correspondence to Raphaël Trouvé, CEBRA, and School of Ecosystem and Forest Sciences, The University of Melbourne, Parkville, Victoria, 3010, Australia; raphael.trouve@unimelb.edu.au.

[1]The following values were originally reported in the literature: NZD\$3.29 billion for New Zealand in 2009, AUD\$13.8 billion for Australia in 2011–2012, USD\$18.9 billion for China in 2004, CAD\$34.5 billion in Canada in 2004, and USD\$100.6 billion for the United States in 2003 and USD\$200 billion in 2010. We first converted each currency to its 2020 value using the inflation calculator of the reserve bank of New Zealand, the reserve bank of Australia, the bank of Canada, and the U.S. bureau of statistics, respectively. We then converted the 2020-NZD\$, AUD\$, and

2009), $11.4 billion for Australia (Hoffmann & Broadhurst, 2016), $26 billion for China (Wan & Yang, 2016), $34 billion in Canada (Colautti, Bailey, Van Overdijk, Amundsen, & MacIsaac, 2006), and $143–238 billion for the United States (Pimentel, Zuniga, & Morrison, 2005; Pimentel, 2011). One option available to the regulator to monitor and reduce biosecurity risk associated with international trade is border inspection, that is, intercepting and stopping pests at point of entry.

The role of border inspection in this context covers two of the three roles identified by Robinson, Burgman, and Cannon (2011), namely, (1) monitoring pathway risk and its evolution to make informed decisions about the threat (e.g., shutting down a risky pathway), and (2) reducing the number of propagules entering the country (sometimes called leakage or slippage, i.e., trade volume $\times$ infestation rate $\times$ leakage rate) or the total pest risk (i.e., volume $\times$ infestation rate $\times$ leakage rate $\times$ impact) arriving in the country by filtering highly infested consignments. In this setting, infestation rate refers to the proportion of units within a consignment that is a biosecurity risk material. Intelligence about the infestation rate of pathways that have the potential to carry quarantine pests is critical to ensure a prompt response to outbreaks and to make defensible decisions about the allocation of inspection effort (Robinson, Chisholm, Mudford, & Maillardet, 2016).

The strategic goals of monitoring and filtering have most often been implemented as an acceptance sampling problem. In acceptance sampling for border biosecurity, the typical recommendation is to inspect all incoming consignments with a sample size n high enough to be "reasonably sure" that the proportion of infested units in each consignment (i.e., the infestation rate p) is below a certain threshold that is deemed "acceptable" by the regulator (IPPC, 2008). Consignments with zero infested samples are deemed compliant, while consignments with at least one infested units are filtered from the system. Using the number of infested units found k and sample size n of the inspection, it is straightforward to estimate the infestation rate of a consignment (up to some margin of error). By collating data from many consignments, we can estimate the mean infestation rate and quantify the threat posed by the pathway.

However, the number of infested units found is seldom reported in regulatory databases. This is a problem. An information that is more often available to the analyst is the proportion or number of noncompliant consignments in the pathway. Since consignments that failed inspection have at least one infested unit (but we do not know the exact number) out of the number of inspected samples, this can be considered as a type of right-censoring of binomial data.[2] It is theoretically possible to recover the mean infestation rate of the consignment from the proportion of consignments that failed inspection and the inspection sampled size (as in Chen, Epanchin-Niell, & Haight, 2018). In our article, we will show how to streamline this estimation procedure by reparameterizing the problem in the framework of generalized linear model (GLM) to leverage existing software solutions and facilitate model extensions. By combining theory and simulated and real inspection data sets, our article also demonstrates that, in the presence of overdispersion (i.e., varying infestation rates among different consignments of the same pathway), using censored data to estimate the mean infestation rate of the pathway will underestimate the threat. Finally, we show that overdispersion can be common in inspection data and that reconstructing pathway mean infestation rate from censored data can lead to a substantial bias in real data sets.

The goals of this article are threefold:

1. Streamline estimating mean infestation rate of a pathway from its proportion of consignments that failed inspection by recasting the problem in the framework of GLMs.
2. Test our ability to reconstruct mean infestation rate from censored measurements using simulated pathways with varying levels of mean infestation rate and overdispersion and using real import pathways.
3. Show that our reconstructed estimates are biased when data are overdispersed and show how prevalent is overdispersion in real biosecurity inspection data sets.

2. MATERIAL AND METHODS

2.1. Acceptance Sampling for Border Inspection

In acceptance sampling for biosecurity border inspection, the typical recommendation is to inspect all incoming consignments with a sample size high

CAD$ to 2020 USD$ using the U.S. federal reserve bank average exchange rates for 2020.

[2]Right censoring corresponds to situations where a data point is above a certain threshold value but it is unknown by how much.

enough to be "reasonably sure" that the proportion of infested units in each consignment (i.e., the infestation rate p) is below a certain threshold that is deemed "acceptable" by the regulator (IPPC, 2008).

Given an infestation rate p, the probability of compliance after inspecting one unit randomly sampled from the consignment is $1 - p$. The probability of compliance after inspecting n units is $(1 - p)^n$ and the probability of noncompliance S (i.e., the consignment-level failure rate or sensitivity of the inspection) follows:

$$S = 1 - (1 - p)^n. \tag{1}$$

This is the basis for the "600 samples" rule often used in biosecurity (Venette, Moon & Hutchison, 2002; IPPC, 2008). Under the "600 samples" rule, consignments are deemed compliant if zero out of 600 randomly sampled units within the consignment are infected. According to Equation (1), the sensitivity (sometimes called the confidence-level of the inspection) for a 600-inspection sample and an underlying infestation rate of 0.5% is $S = 1 - (1 - 0.005)^{600} \sim 0.95$: in the long run, the "600 samples" rule will detect ~95% of the consignments having an infestation rate of 0.5%.

2.2. Estimating the Mean Infestation Rate of a Pathway from Censored Measurements

If we know the proportion of consignments that failed inspection S in a pathway and the number of units inspected per consignment n, we can solve for the infestation rate p:

$$p = 1 - (1 - S)^{1/n}. \tag{2}$$

This equation is again best illustrated using the "600 samples" rule: a pathway in which 95 out of 100 consignments failed a 600-sample inspection should have a mean infestation rate $p = 1 - (1 - 95/100)^{1/600} = 0.005$.

There are three issues in using Equation (1) to reconstruct the infestation rate from censored measurements: First, the proportion of failed consignments is only known from discrete data (i.e., we do not directly observe the proportion of failed consignments, but instead the number of failed consignments out of a number of inspected consignments). To account for this binomial stochasticity, we can assume that the total number of failed consignments in the pathway comes from a binomial distribution (or equivalently, we can assume that each consignment is a 0/1 Bernoulli experiment, see Chen et al., 2018). Second, p is bounded in the 0–1 range, which is sometimes problematic for the fitting algorithms, especially when we want to quantify parameter uncertainties. Both these issues can be by-passed by reparameterizing Equation (1) in the framework of GLMs with a Bernoulli distribution and a complementary log-log ("cloglog") link. We show the equivalence between Equation (1) and its GLM parameterization in Appendix A. The third and main issue, which we will ignore in this section but cover at length in the rest of the article, arises when p is not constant—but varies among different consignments of the same pathway (i.e., when pathways are overdispersed). This will cause bias in the mean infestation rate per pathway estimated from censored-measurement data.

But first, for pedagogical purposes, we will ignore overdispersion. In the GLM model for censored measurements, the noncompliance status of each consignment (zero if compliant, one if noncompliant) is modeled as being sampled from a Bernoulli distribution with mean given by the cloglog version of the sensitivity equation (see Appendix A for the equivalence between Equations (1) and (3)). The model for censored-measurement data follows:

$$\begin{aligned} k^* &\sim Bernoulli(S) \\ S &= 1 - exp(-exp(\alpha + log(n))), \end{aligned} \tag{3}$$

where k^* is a right-censored version of k, a vector of the number of infested units found in each inspection (i.e., k^* equals zero if $k = 0$ and k^* equals one if $k \geq 1$, essentially reporting the compliance or noncompliance status of the consignment), S represents the sensitivity of each inspection (i.e., the probability of each consignment being noncompliant), α is the mean infestation rate of the pathway on the cloglog scale (see Appendix A), and n is a vector representing the sample size of each inspection. We obtain the mean infestation rate p_{cens}[3] reconstructed from censored-measurement data by applying the inverse-cloglog function: $p_{cens} = 1 - exp(-exp(\alpha))$ (i.e., we fix $n = 1$ in Equation (3)).

For example, if 95 consignments failed compliance and 5 are deemed compliant in a pathway, we can estimate the mean infestation rate of the

[3]Here, we use the p_{cens} parameter to indicate that the estimate comes from censored-measurement data and to distinguish it from estimates computed from detailed and more direct data on the number of infested units k found in each inspection.

pathway by fitting this vector of binary consignment-failure data using a GLM with a Bernoulli error term and a cloglog link, adding an exposure term (sometimes called offset) of log(600) to account for the number of inspected units per inspection.[4] The procedure gives an estimate for the intercept $\alpha = -5.3$, which translates to a mean infestation rate p_{cens} of $1 - exp(-exp(-5.3)) = 0.005$. We recover the proportion of noncompliant consignments (i.e., the sensitivity of the inspection) by adding the 600 samples per consignment exposure: $S = 1 - exp(-exp(-5.3 + log(600))) = 0.95$.

2.3. Estimating the Mean Infestation Rate of a Pathway from Detailed Data on the Sample Size *n* and the Number of Infested Units *k* per Inspection

When we have access to detailed data on the sample size n and number of infested units k per inspection, we can directly estimate the mean infestation rate p of each consignment or of the whole pathway. There are different ways to estimate this quantity (Brown, Cai, & DasGupta, 2001), but for the purpose of comparing our results with the censored-measurements method (Equation (3)), we will also use a GLM model with a "cloglog" link. The model for uncensored data follows:

$$\begin{aligned} k &\sim Binom(p, n), \\ p &= 1 - exp(-exp(\alpha)), \end{aligned} \tag{4}$$

where k is the observed number of infested units per inspection, n in the sample size of the inspection, p is the mean infestation rate, and α is the infestation rate on the cloglog scale. Model inputs k and n are single integers when we estimate p for a single consignment but are vector integers when we estimate p for a whole pathway (i.e., several consignments).[5]

[4] Using the R software (R, 2018), the syntax for fitting Equation (3) to such a vector of binary consignment-failure data is: k_star = c(rep(1, 95), rep(0, 5)); n = rep(600, 100); glm(k_star ~ 1 + offset(log(n)), family = binomial(link = "cloglog")).

[5] In R, the syntax for fitting Equation (4) to, e.g., three inspected consignments with zero, two, and five infested units in a 600-samples inspection is: k = c(0, 2, 5); n = c(600, 600, 600); glm(cbind(k, n-k) ~ 1, family = binomial(link='cloglog')).

2.3.1. *Estimating the overdispersion parameter σ of the pathway*

When we have access to detailed data on the sample size n and number of infested units k per inspection, we can allow for varying infestation rate among consignments of the same pathway (i.e., overdispersion) by adding a consignment-level random effect (varying intercept) terms to Equation (4) (Harrison, 2014).

$$\begin{aligned} k_j &\sim Binom(p_j, n_j), \\ p_j &= 1 - exp(-exp(\alpha_j)), \\ \alpha_j &\sim Normal(\alpha, \sigma). \end{aligned} \tag{5}$$

Due to the symmetry of the normal distribution, the estimated population parameter α represents both the mean and the median of the α_j distribution on the cloglog scale. However, in the original scale, the distribution of p_j is typically asymmetric and right-skewed because the p_j values for most of the consignments are close to zero and the presence of a few consignments with a higher infestation rate creates a heavy right tail.

We show in Appendix B that the inverse-cloglog transform of α: $\tilde{p} = 1 - exp(-exp(\alpha))$ represents the median of the distribution of p_j—rather than its mean. In a right-skewed distribution, the median of the distribution $\tilde{p} = 1 - exp(-exp(\alpha))$ is lower or equal to the mean $p = \overline{p_j} = \overline{1 - exp(-exp(\alpha_j))}$. The mean infestation rate p can be estimated using Monte Carlo simulations: we first sample several α_j values from a normal distribution with mean α and standard deviation σ. We then compute the p_j values using the inverse cloglog function and compute $p = \overline{p_j}$ as our estimate of the mean infestation rate. Alternatively, the mean infestation rate p of the p_j distribution can be approximated by using the median of the distribution and a correction term for the bias derived using Taylor expansion for the moments of functions of random variables (Appendix B).

It is worthwhile noting that the asymmetry of the p_j distribution is not an artifact of using a cloglog link to model the overdispersion, but stems from the boundedness of the p_j parameters and the fact that a majority of the values are clustered on the lower part of the 0–1 range: If the median infestation rate is already quite low, then there is not a lot of room left for the consignments-specific p_j to be much lower, but there is a lot of room for p_j values to increase. Alternative ways to model overdispersion on data bounded to the 0–1 range (consignment-level

random effect based on a GLM with a logit link, beta-binomial distribution, ...) would behave similarly.

2.4. Evaluating Our Reconstruction of Mean Infestation Rate Estimated from Censored Measurement

2.4.1. Simulating Pathways with Varying Dgree of Overdispersion

To check if we can recover the mean infestation rate of a pathway using only its consignment-level failure rate, we ran a computational experiment where we know the true mean infestation rates of the pathway. Since we suspect that overdispersion might affect our result, we simulated pathways with varying degree of overdispersion. We then evaluated our method by comparing the estimated mean infestation rate per pathway from censored measurements (using Equation (3)) with the true simulated infestation rate.

We fixed the design of our computational experiment by simulating pathways with different values of the infestation rate parameter on the cloglog scale ($\alpha = -9.2, -8.1, -6.9, -5.8, -4.6, -3.5, -2.3$, and -1.0, i.e., median infestation rates of $\sim 10^{-4}$, 3×10^{-4}, 10^{-3}, 3×10^{-3}, 0.01, 0.03, 0.1, 0.3) and with overdispersion parameter ($\sigma = 0$, 0.5, 1, and 2). These values where selected to cover a large range of potential mean infestation rates and overdispersion found in typical biosecurity pathways. We simulated 40 pathways corresponding to the 40 possible combinations of α and σ.

For each of these 40 simulated pathways, we followed the three steps below:

1. We simulate 100 consignments with varying infestation rate per consignment. We first sampled 100 varying α_j values on the cloglog scale from a normal distribution with mean α and standard deviation σ. These 100 consignment-specific values were then back-transformed to the infestation rate scale by using an inverse cloglog function ($p_j = 1 - exp(-exp(\alpha_j))$). At this stage, we estimate the true simulated mean p of the pathway as the mean of these 100 p_j values.
2. We simulate a 600-unit inspection for each of the 100 consignments of the pathway. The compliance status of each consignment is sampled from a Bernoulli distribution with mean given by the sensitivity of the inspection (Equation (1), using the consignment-specific p_j and $n = 600$). This gives us the number of noncompliant consignments out of 100.
3. We then estimate the mean infestation rate of the pathway p_{cens} by fitting Equation (3) to the number of noncompliant consignments out of 100 consignments.

For each of the 40 simulated pathways, we repeated the procedure 1,000 times and summarized the results by using the mean and 95% percentile range of the 1000 mean infestation rate estimates.

2.4.2. Case Study with Real Border Inspection Data

To assess whether we can reliably reconstruct the infestation rate of pathways from consignment-level failure data and to quantify the degree of overdispersion typical in real import pathways, we used a database of live germplasm interception data to Australia. This specific data set was chosen because each inspection reported the detailed sample size n of the inspection and the number of infested units k in the sample. From these detailed data, we then computed a censored data set reporting only the number of noncompliant consignments in each pathway (i.e., consignments were considered to fail inspection when at least one infested unit was found during the inspection).

In this germplasm pathway data set, we filtered the data to only retain inspections with 20–1000 inspected units. When then filtered the data to only retain genus that had at least 30 consignments. We were left with 9,364 consignments, 55 genus, from 184 importers across 26 countries. The median number of inspected units per consignment was 77 (2.5–97.5% percentile range of 22–630) and the mean infestation rate per consignment was of 1.4% (2.5–97.5% percentile range of 0–14%). In this data set, we considered each genus to be an individual pathway.[6]

We computed the mean infestation rate p of each pathway using the detailed data on the number of infested units and the number of sampled units per pathway using Equation (4). We also estimated the mean infestation rate p_{cens} of the pathway from its consignment-level failure data using Equation (3).

[6]While further stratification of the pathway is possible in theory (e.g., genus × importer), there is typically not enough consignments per stratum (e.g., if we were to stratify one of our largest pathway (528 consignments) by importer companies (83 importers), only 20% of the importers would have >10 inspected consignments) to get reliable estimates of mean infestation rate.

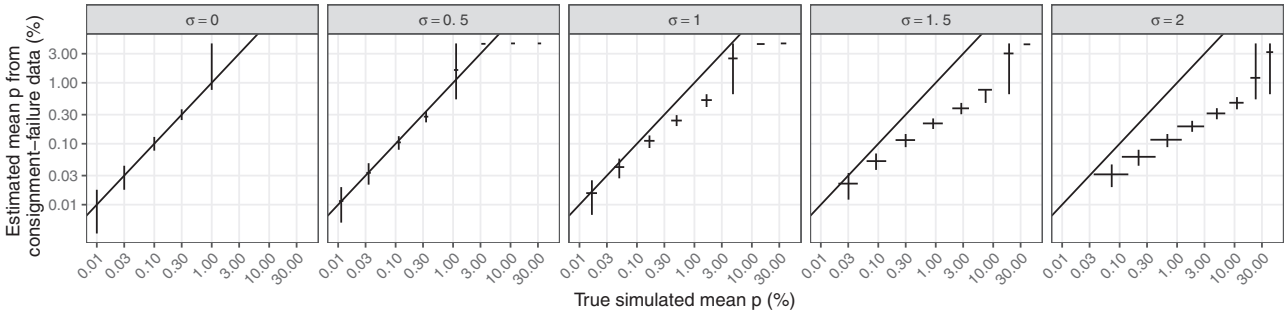

Fig 1. Estimated infestation rate using consignment-level failure (censored-measurements) vs. true simulated infestation rate. Dots indicate the mean of 1,000 replicates of the simulation. The bars indicate 5–95% quantiles of the replicated simulations. The solid lines show the 1:1 line (estimated p_{cens} = true p). Dots below the solid line show that p_{cens} underestimates the true p.

Additionally, to visualize whether eventual differences in the estimates of p_{cens} versus p can be attributed to pathway overdispersion, we estimated σ for each pathway by fitting Equation (5) to the detailed data on the number of infested units and the number of sampled units per pathway, allowing for a consignment level random effect on the intercept. To further check whether the differences in the estimates of p_{cens} versus p can be explained by pathway-specific overdispersions, we used Equation (B1) to compute a bias-corrected p_{cens} that corrects for the overdispersion of the pathway (Appendix B).

3. RESULTS

3.1. Simulated Pathways with Increasing Overdispersion

In the absence of overdispersion (left panel of Fig. 1) and when true p is <3%, p_{cens} estimated from consignment-failure data matches the true simulated p (the dots follow the 1:1 line). However, when the true infestation rate is above a certain threshold ($\geq$ 3%), estimated p_{cens} saturates leading to an underestimation bias. This threshold happens when the true infestation rate is too high and 100% of the consignments are noncompliant: The estimated mean p_{cens} plateaus when p_{cens} is high enough to predict close to 100% consignment failure. For a 600-unit inspection, this plateau happens for an infestation rate of 3% (i.e., the sensitivity of the inspection in Equation (1), $1-(1-0.03)^{600}$ is nearly indistinguishable from one). Working with censored measurements and a GLM approach will typically underestimate the infestation rate when the true infestation rate is too high. Note that when using the GLM approach, the specific value estimated for p_{cens} when 100% of the consignments are noncompliant will be sensitive to the optimization algorithm used, the tolerance parameter that defines when we have reached convergence, and the starting parameter values of the optimization procedure. By contrast, when 100% of the consignments are noncompliant, the simple formula given in Equation (2) will instead predict $p_{cens} = 100\%$ (i.e., overestimate the infestation rate).

As we increase the overdispersion in our simulated data (i.e., the variability in the infestation rates among consignments within each pathway increases, see Panels 2–5 in Fig. 1), the estimated p_{cens} from consignment-level failure data show increasing negative bias. This underestimation of true p by p_{cens} also increases with the value of the true infestation rate. This is most visible in the right panel of Fig. 1 (high overdispersion with $\sigma = 2$), in which the $p_{cens}/p = 0.1$ when true p is 0.1% (i.e., a 10-fold underestimation), but the bias becomes $p_{cens}/p = 0.04$ when true p = 10% (i.e., a 40-fold underestimation).

3.2. Case Study in the Germplasm Data Pathways

With the exception of two pathways that have a low mean infestation rate and overdispersion at the bottom left of Fig. 2(a), there was a systematic downward bias in the p estimated from censored-measurements in our germplasm interception data set (Fig. 2(a)). Mean infestation rate estimated from censored measurements were an average ~ 0.3 times lower than expected, with the downward bias increasing with increasing values of the overdispersion parameter of the pathway σ.

As already noted, the origin of the bias is well understood and, if we know the $\overline{\alpha}$ and the overdispersion parameter σ of the pathway, then we can correct for the bias using Taylor expansion method (Equation (B1), Fig. 2(b)). The bias-correction term

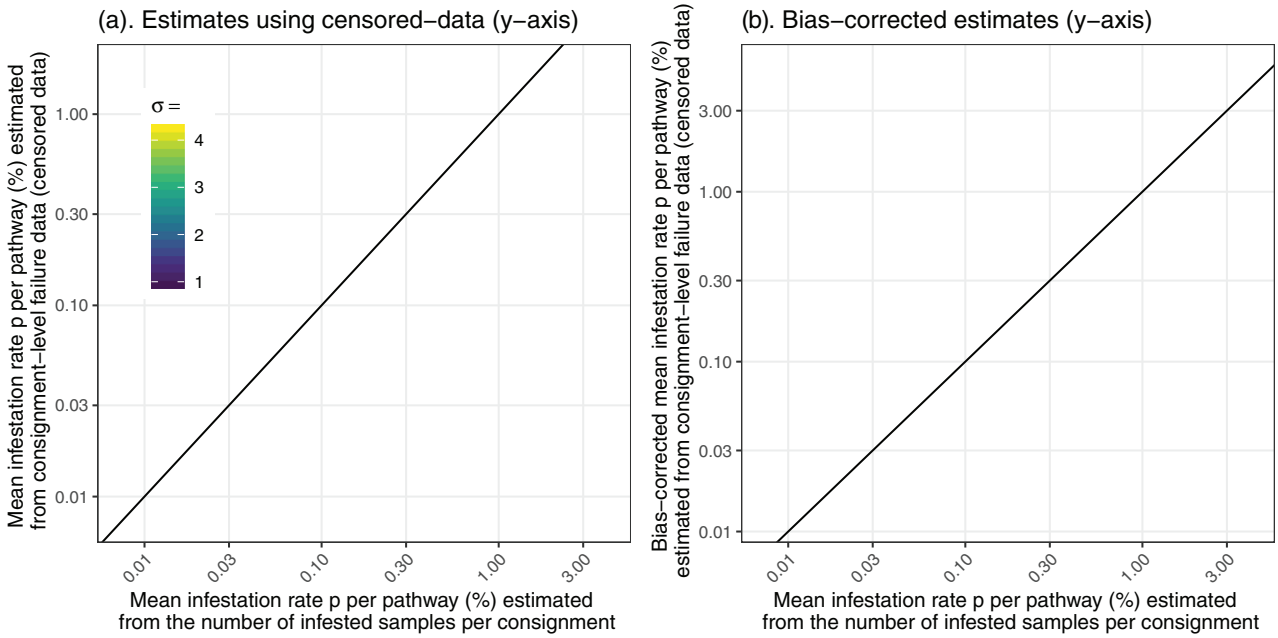

Fig 2. Comparison of the mean infestation rate per pathway estimated using the detailed data (inspection sample size n and the number of infested units found k in each consignment) vs. the censored measurements (number of failed consignments and total number of inspections). Each dot represents the mean infestation rate of the genus estimated from the full data (x-axis) and the censored measurements (y-axis). (b.) Same analysis but applying the bias correction from Equation (B1) to the estimates obtained from the censored-measurements. The σ value of each pathway used for the bias correction was estimated by fitting Equation (5) to the detailed data. The solid lines show the 1:1 line (estimated p = true p). Points below the line show that the estimated p underestimates the true p.

is a nonlinear function of $\overline{\alpha}$ and σ (Fig. B1). Unfortunately, even if we understand where the bias comes from, we cannot correct for the bias in practice as we cannot estimate the overdispersion parameter σ from the binary censored measurements. Sadly, if we only have access to censored measurements, we are not able to know if there is a bias.

Yet, using the detailed data on the number of infested units per inspection, we found evidence for overdispersion in 38 of the 40 germplasm import pathways: with the exception of two pathways at the bottom left of Fig. 2, model fits were systematically better (in terms of widely applicable information criteria, WAIC; Watanabe, 2013) with overdispersion (Equation (5)) than without overdispersion (Equation (4)). While this shows that we can reliably reconstruct the mean infestation rate from censored measurements in the absence of overdispersion (e.g., the two pathways without overdispersion at the bottom left of Fig. 2(a). sit right on the 1:1 line), it also shows that overdispersion is the rule rather than the exception in our data set.

Finally, note that the overdispersion parameter cannot be estimated from the Bernoulli distributed censored measurements. The bias correction in Fig. 2(b) is essentially an exercise to see if we understand the source of the bias. It is not a bias correction that we will be able to apply in practice.

4. DISCUSSION

In many applications of biosecurity, it can be useful to know the mean infestation rate of different import pathways (e.g., to quantify the import risk and prioritize inspection effort). However, we do not always have access to detailed data on the number of infested units found in each inspected consignment. Rather, we more often have access to the proportion or number of consignments that were rejected during the inspection (i.e., consignments with $\geq$ one infested sample in the inspection). While it is tempting to use this censored measurements to reconstruct mean infestation rate of the pathway, this will typically underestimate the true infestation rate (by an average factor of $\sim$ 0.3 in our germplasm data set). This underestimation issue is due to overdispersion (i.e., varying infestation rate among different consignments of the same pathway) and the issue increases with increasing levels of overdispersion and median infestation rate of the pathway.

4.1. Reconstructing Infestation Rate from Censored-Measurements in the Absence of Overdispersion

When there is no overdispersion, it is possible to reconstruct the mean infestation rate of a pathway if we know the number of consignments that failed inspection (i.e., inspection with ≥ one infested unit) and the number of inspected units per consignments (left panel of Fig. 1 with $\sigma = 0$).

However, reconstructing the mean infestation rate from censored measurements fails for pathways with high infestation rate. Specifically, the reconstructed p_{cens} plateaus when 100% of the consignments fail compliance: in such case p_{cens} is (wrongly) estimated to be the minimum infestation rate that predicts close to 100% of the consignment to fail compliance. For example, for a 600 units inspection and true infestation rates ≥3%, p_{cens} values saturate at 3% (left panel of Fig. 1), as the sensitivity $1-(1-0.03)^{600}$ (Equation (1)) is almost indistinguishable from one. Counterintuitively, increasing sample size will make this saturating bias at high infestation rate more noticeable (e.g., for a 3,000 units inspection, the saturation will happen for true infestation rates ≥0.3%, as $1-(1-0.003)^{3000}$ is almost indistinguishable from one).

Getting data from additional inspections—rather than increasing the sample size of each inspection—might help with this saturation bias. Additional inspections data will increase the resolution on the proportion of consignments that fail inspection and increases the likelihood of having at least one compliant consignment in the pathway, allowing us to get identification. The saturation issue might also be mitigated by pooling several pathways together and modeling them using hierarchical models (Gelman et al., 2014). In hierarchical models, pathways where the infestation rate can be estimated (i.e., pathways that do not have 0% or 100% of consignment failing inspection) will help estimating the infestation rate of pathways that are not well identified.

4.2. Reconstructing Infestation Rate from Censored Measurements in the Presence of Overdispersion

Reconstructing the mean infestation rate from censored measurements also fails when the data are overdispersed. Combining censored measurements with overdispersed data can lead to a substantial (e.g., an order of magnitude) underestimation bias.

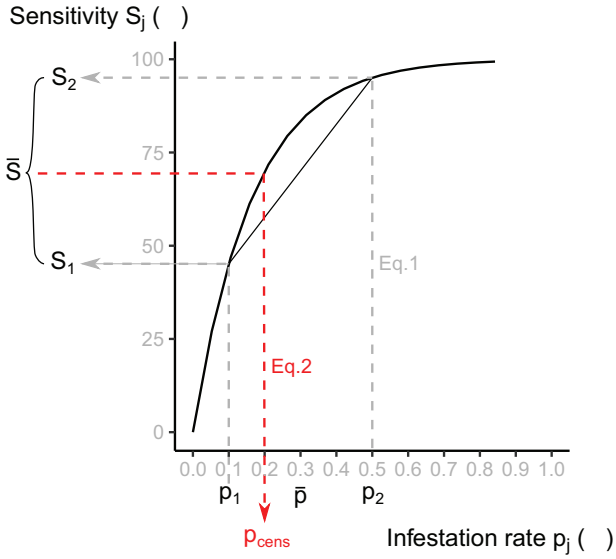

Fig 3. Illustration of the effect of Jensen's inequality on the estimation of infestation rate p_{cens} from the observed proportion of consignment that failed inspection ($\overline{S}$) and in the presence of overdispersion. We used Equation (1) to map from the x-axis to the y-axis and use Equation (2) to map from the y-axis to the x-axis. We have two consignments with infestation rate p_1 and p_2 and mean infestation rate $\overline{p}$. We use Equation (1) to compute the sensitivity S_1 and S_2 of each consignment under a 600-unit inspection (gray arrows). The mean sensitivity, which is also the proportion of failed consignments that we would observe in our empirical data is $\overline{S}$. Now if we were to ignore the overdispersion and use Equation (2) to try reconstructing the mean infestation rate from $\overline{S}$ (red arrow), we obtain p_{cens}. As shown in the figure, p_{cens} underestimates $\overline{p}$. If the points p_1 and p_2 were further apart (i.e., higher overdispersion), the underestimation bias would be higher.

The origin of the bias is well understood and comes from Jensen's inequality, which states that convex transformation of a mean is lower than or equal to the mean applied after convex transformation (i.e., $p_{cens} = f(\overline{S_j}) \leq \overline{f(S_j)} = \overline{p}$) (Fig. 3).

In the presence of overdispersion, each consignment has its own infestation rate p_j and we are interested in estimating the mean infestation rate $\overline{p}$ of the pathway. When using censored measurements, we do not observe individual p_j's but the proportion of consignments that failed inspection $\overline{S}$. Since p_j and S_j are linked with the sensitivity equation (Equation (2)), we can try reconstructing $\overline{p}$ from $\overline{S}$. However, in the presence of overdispersion, the method fails due to the convex nature of the function allowing us to calculate p from S (be it Equation (2) or Equation (3)).

The issue is best illustrated using a concrete example (Fig. 3): Say we have two consignments with infestation rate p_1 and p_2 and mean infestation rate $\overline{p}$. We then use Equation (1) (gray arrows) to

compute the sensitivity S_1 and S_2 associated with each consignment under a 600-unit inspection. The mean sensitivity, i.e., the proportion of failed consignments that we would observe in our empirical data, is $\overline{S}$. Now if we were to ignore overdispersion and directly use Equation (2) (red arrow) to try reconstructing the mean infestation rate from $\overline{S}$, we would obtain p_{cens}, which underestimates $\overline{p}$ due to the convex nature of Equation (2) (Fig. 3):

$$p_{cens} = 1 - (1 - \overline{S_j})^{1/n} \leq \overline{1 - (1 - S_j)^{1/n}} = \overline{p}. \quad (6)$$

Graphically, we can see that the amounts by which p_{cens} underestimates $\overline{p}$ depends on the degree of curvature of the function relating S to p (the higher the curvature, the higher the bias) and the degree of overdispersion of the pathway (the further p_1 and p_2 are apart from each other, the higher the bias) (Fig. 3). Thus, if we know the degree of overdispersion in p and the curvature around the function evaluated around p, we might come up with a correction term for this bias. This is exactly what the Taylor expansion allows us to do (Equation (B1)): The bias correction term depends on the second derivative of our function (i.e., the curvature) and the overdispersion parameter σ.

Note that the fact that the bias increases with overdispersion does not depend on the specifics of the model used to reconstruct the mean infestation rate of a pathway from censored measurements (i.e., whether we are using Equation (2) or Equation (3)). Rather, the presence of the bias is an unavoidable consequence of the convex nature of the function relating S to p and the bias expresses itself only in the presence of overdispersion.

Note that while we understand the origin of the bias well (Appendix B) and can correct for it when we know the overdispersion parameter σ of the pathway (Fig. 2(b)), in practice we cannot correct for the bias as we cannot estimate the overdispersion parameter σ from Bernoulli distributed censored measurements data. Worse, if we only have access to censored measurements, we cannot know if there is overdispersion and thus if there is a bias.

4.2.1. Overdispersion Seems to Be Common in Biosecurity Inspection Data

In our germplasm import data set, 38 of the 40 pathways show evidence of overdispersion (i.e., all the pathways that are below the 1:1 line in Fig. 2(a)). Overdispersion is also present in other published biosecurity inspection data sets (e.g., the Kiwi import pathway to Japan, fig. 2 of Yamamura & Sugimoto, 1995).

The cause of overdispersion in biosecurity inspection data is diverse. For one, no consignment (even from the same exporter company) is exactly the same as the other ones. Product quality and the exposure to different pests and diseases might change with time and even the most rigorous quality management process will allow for some variation. Additionally, it is common for pests to have aggregated distributions (Hughes & Madden, 1992). Aggregated distributions will lead pests to cluster preferentially in certain consignments and not in others. This initial variability in infestation rate might further be amplified by exponential pest growth during transit from the exporter to importer country. Part of the overdispersion observed in biosecurity data also likely comes from a lack of data stratification on the part of the analyst (e.g., we stratified pathways by genus, but we might also have considered stratifying by country of origin, importer companies, ...). However, we caution that additional stratification can leads to strata with a number of consignments too low to make reliable inference. For example, if we were to stratify one of our largest pathway (528 consignments) by importer companies (83 importers), only 20% of the importers would have >10 inspected consignments. There would not be enough consignments per strata to get reliable estimates of mean infestation rate. Furthermore, when we tried explaining some of the variability in infestation rate among consignments by adding covariates (e.g., country of origin, importer and exporter companies, year, and month of inspection, with an interaction between year and month of inspection) to Equation (5), we were only able to reduce the overdispersion parameter σ per pathway by an average of 14% compared to using Equation (5) without covariates. This shows that a large part of the variability in infestation rate among consignments is irreducible given the level of information typically available to biosecurity analysts.

Given that overdispersion seems to be common in biosecurity data, if we cannot estimate the overdispersion of the pathway because we only have access to censored data, we suggest being conservative and assuming that overdispersion is present. This means that in most cases, we should avoid estimating the mean infestation rate of a pathway from censored-data as it will underestimate the true infestation rate of the pathway.

4.2.2. *Consequences of Using Biased Estimates of Mean Infestation Rate for Monitoring and Targeted Sampling*

Monitoring pathway infestation rate using censored data will likely lead to underestimating the biosecurity risk associated with different pathways. This might give a false sense of confidence and impair making informed decisions about the threat.

While shifting sampling effort from low-risk to high-risk pathways (i.e., targeted sampling) has been shown to reduce leakage (by a factor of 0.7–0.8; Robinson et al., 2011; Springborn, Lindsay & Epanchin-Niell, 2016; Chen et al., 2018), targeted sampling based on biased estimates of infestation rates can potentially be harmful (see also Powell, 2015). Since overdispersion (and thus bias) will vary among pathways, targeted sampling will likely be sub-optimal, and might even be detrimental in cases where we decrease sampling effort in pathways that were thought to have low infestation rate (but had high infestation rate) to redistribute it to other pathways.

For example, we might imagine two pathways with different mean infestation rate and overdispersion: Pathway A has a true infestation rate $\sim 1\%$, estimated to be $\sim 0.1\%$ due to high overdispersion. Pathway B has a true infestation rate of 0.3%, estimated to be $\sim 0.3\%$ as it has low overdispersion (these two types of pathways can be found in Fig. 2(a)). In such a case, targeted sampling will recommend increasing inspection sample size on pathway B and decreasing on pathway A. Unfortunately, it is the opposite of what should be done if we wanted to reduce the leakage.

4.2.3. *Importance of compiling all the data*

Reconstructing the mean infestation rate of pathways from censored measurements leads to biased estimates. To counteract this issue, we suggest collecting and reporting detailed data on the number of infested units and the sample size of each inspected consignment instead of only its compliance status. This will support reliable estimates of pathway mean infestation rate.

Additionally, the detailed data will also allow estimating the overdispersion parameter of each pathway. Knowing the degree of overdispersion of different pathways is important. We might want to treat pathways with similar mean infestation rate but different overdispersion differently. For example, when there is high overdispersion in a pathway, inspection act as a sieve, preferentially filtering highly infested consignments and accepting consignments with a low infestation rate. With high overdispersion, even small sample size inspections will weed out the few highly infested consignments, greatly reducing the mean infestation rate in accepted consignments. Also, while in the absence of overdispersion, the optimization procedure for targeted sampling often suggest investing maximum sampling effort in the most infested pathways and not sampling lower infested pathways (Chen et al., 2018), in the presence of overdispersion, we would likely want to inspect most pathways with at least a small sample size to weed out the few highly infested consignments.

On the other hand, if all the consignments of a pathways have similar infestation rate (no overdispersion), there will be little improvement to be made in reducing the infestation rate before and after inspection. Worse, in the absence of overdispersion, the inspection will randomly filter some of the consignments according to the sensitivity of the test, but the infestation rate of these noncompliant consignments will be no different than that of the accepted consignments. While this is an extreme case, it highlights that we often implicitly assume that there is some degree of overdispersion in the pathways: We assume that infestation rate in accepted consignments is lower that in the filtered consignment. Otherwise there would be no point in doing inspections. If most pathways are overdispersed and if the degree of overdispersion affects inspection efficiency, we suggest there is valuable intelligence to be gained in estimating pathway overdispersion.

5. CONCLUSION

While in theory it is possible to leverage Equation (1) (or its reparameterization Equation (3)) to reconstruct the mean infestation rate of different pathway from censored measurements (i.e., the proportion of consignment that failed in the pathway), the methods greatly underestimate mean infestation rate when there is overdispersion (i.e., varying infestation rate among different consignments of the same pathway). Since overdispersion is common in biosecurity data and since the consignment-level failure data cannot be used to detect overdispersion, we advise against reconstructing infestation rate from censored measurements. Instead, we recommend recording detailed data (the sample size n

and the number of infested units k) of all inspected consignments.

ACKNOWLEDGMENTS

The authors would like to thank the Department of Agriculture, Water and the Environment of Australia for providing case study border inspection data. This research was supported by the Centre of Excellence for Biosecurity Risk Analysis, CEBRA. We thank three anonymous reviewers for their constructive comments, which helped improve the article. We also thank Mei Bai for providing feedback on an early draft of the article.

APPENDIX A

In this appendix, we show the equivalence between Equation (1) and the cloglog parameterization used in a GLM framework.

We begin with Equation (1)

$$S = 1 - (1-p)^n,$$
$$S = 1 - exp(log(1-p) \times n). \quad \text{(A1)}$$

Since for $p \in (0,1)$,

$$log(1-p) = -exp(log(-log(1-p))).$$

Then, $log(1-p) \times n = -exp(log(-log(1-p))) \times n,$

$$log(1-p) \times n = -exp(log(-log(1-p)) + log(n)). \quad \text{(A2)}$$

Replacing $log(1-p) \times n$ in Equation (A1) by $-exp(log(-log(1-p)) + log(n))$ in Equation (A2) gives the following expression for the sensitivity:

$$S = 1 - exp\left(-exp\left(\underbrace{log(-log(1-p))}_{\text{cloglog link}=\alpha} + \underbrace{log(n)}_{\text{log(exposure)}}\right)\right).$$

The infestation rate p after inspecting one unit can be computed as:

$$log(-log(1-p)) = \alpha,$$
$$-log(1-p) = exp(\alpha),$$
$$log(1-p) = -exp(\alpha),$$
$$p = 1 - exp(-exp(\alpha)).$$

The last equation is the inverse–cloglog link function typically found in textbooks on complementary log–log regression. The inverse–cloglog function effectively bounds the α parameter from the $-\inf + \inf$ range to the 0–1 range and allows streamlining estimations of p using standard GLM methods.

We obtain the probability of noncompliance after inspecting n units by adding the exposure term for the number of inspected units n:

$$S = 1 - exp(-exp(\alpha + log(n))).$$

Since the data itself are binary (compliant consignment = 0, noncompliant consignment = 1), it is appropriate to use a Bernoulli error distribution to fit the data, leading to Equation (3).

APPENDIX B

In this appendix, we derive a bias-correction term for the inverse–cloglog function in the presence of overdispersion.

Using Taylor expansion for the moments of functions of random variables (sometimes called the a second-order "delta" method), the mean of a function f of a random variable X can be approximated from the function evaluated at the mean of X and a bias correction term. The bias correction terms is the product between the second derivative of the

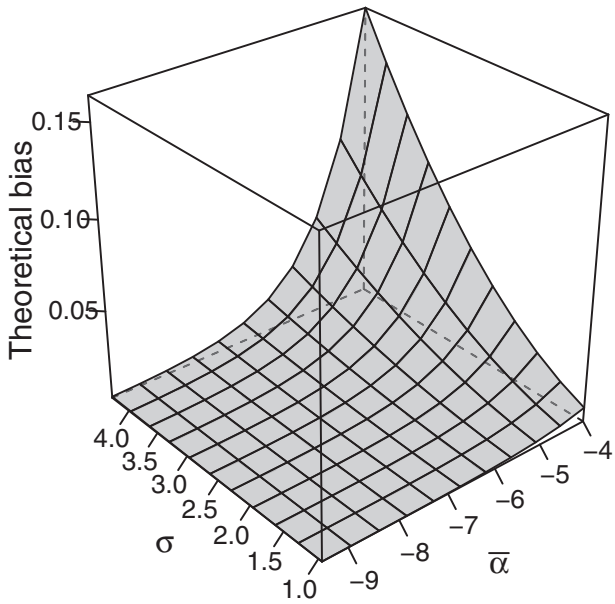

Fig B1. Theoretical bias correction term for the mean infestation rate vs. pathway overdispersion (σ) and median infestation rate ($\overline{\alpha}$, expressed in the cloglog scale) computed from Equation (B1).

function f evaluated at the mean of X and the variance of the random variable σ^2. The general formula follows:

$$\overline{f(X)} \approx f(\overline{X}) + f''(\overline{X})\frac{\sigma_X^2}{2}.$$

Our function of interest is the inverse–cloglog developed in Equation (3). The function and its second derivative follow:

$$f(\alpha_j) = 1 - exp(-exp(\alpha_j)),$$
$$f''(\alpha_j) = -exp(\alpha_j - exp(\alpha_j))(exp(\alpha_j) - 1),$$

where α_j a random variable sampled from a normal distribution with mean $\overline{\alpha}$ and standard deviation σ. Thus, the bias-corrected mean infestation rate $p = \overline{f(\alpha_j)}$ of a pathway can be approximated following:

$$p = \overline{1 - exp(-exp(\alpha_j))} \approx \overbrace{1 - exp(-exp(\overline{\alpha}))}^{\text{Median of the } p_j \text{ distribution}} \underbrace{-exp(\overline{\alpha} - exp(\overline{\alpha}))(exp(\overline{\alpha}) - 1))\frac{\sigma^2}{2}}_{\text{Bias correction term}}. \quad \text{(B1)}$$

The bias-correction terms in Equation (B1) increases with the overdispersion terms σ and with the $\bar{\alpha}$ parameter (i.e., median infestation rate expressed in the cloglog scale) (Fig. B1).

Demonstration that $1 - exp(-exp(\overline{\alpha}))$ is the median of the p_j distribution. Since the distribution of α_j is Gaussian and symmetric, 50% of the mass of the distribution is below $\overline{\alpha}$ and 50% is above it: $\overline{\alpha}$ is both the mean and the median of the α_j distribution. While the inverse–cloglog transformation that maps α_j to p_j stretches the α_j values on both sides of $\overline{\alpha}$ differently, the inverse–cloglog transformation is monotone: α_j values that are below $\overline{\alpha}$ map to p_j values below $1 - exp(-exp(\overline{\alpha}))$ and α_j values that are above $\overline{\alpha}$ map to p_j values above $1 - exp(-exp(\overline{\alpha}))$. This means that we have 50% of the mass of the p_j distribution below $1 - exp(-exp(\overline{\alpha}))$ and 50% of the mass above it: The inverse–cloglog function of $(\overline{\alpha})$ is the median of the p_j distribution.

REFERENCES

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101–133.

Chen, C., Epanchin-Niell, R. S., & Haight, R. G. (2018). Optimal inspection of imports to prevent invasive pest introduction. *Risk Analysis*, *38*(3), 603–619.

Colautti, R., Bailey, S., Van Overdijk, C., Amundsen, K., & MacIsaac, H. (2006). Characterised and projected costs of nonindigenous species in canada. *Biological Invasions*, *8*, 45–59.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall.

Giera, N., & Bell, B. (2009). Economic costs of pests to New Zealand. MAF Biosecurity New Zealand Technical Paper 2009/31, Biosecurity New Zealand, Ministry of Agriculture and Forestry, Wellington.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, *2*, e616.

Hoffmann, B. D., & Broadhurst, L. M. (2016). The economic cost of managing invasive species in Australia. *NeoBiota*, *31*, 1–18.

Hughes, G., & Madden, L. (1992). Aggregation and incidence of disease. *Plant Pathology*, *41*(6), 657–660.

IPPC (2008). International standard for phytosanitary measures. ISPM No. 31: Methodology for sampling of consignment. Technical report, Secretariat of the International Plant Protection Convention.

Pimentel, D. (2011). *Biological invasions: Economic and environmental costs of alien plant, animal, and microbe species* (2nd ed.). Hoboken, NJ: CRC Press.

Pimentel, D., Zuniga, R., & Morrison, D. (2005). Update on the environmental and economic costs associated with alien-invasive species in the United States. *Ecological Economics*, *52*, 273–288.

Powell, M. R. (2015). Risk-based sampling: I don't want to weight in vain. *Risk Analysis*, *35*(12), 2172–2182.

R, D. C. T. (2018). *R: A language and environment for statistical computing. Version 3.5.1*. R Foundation for Statistical Computing. http://www.R-project.org.

Robinson, A., Burgman, M. A., & Cannon, R. (2011). Allocating surveillance resources to reduce ecological invasions: Maximizing detections and information about the threat. *Ecological Applications*, *21*(4), 1410–1417.

Robinson, A., Chisholm, M., Mudford, R., & Maillardet, R. (2016). Ad hoc solutions to estimating pathway non-compliance rates using imperfect and incomplete information. In F. Jarrad, S. Low-Choy, & K. Mengersen (Eds.), *Biosecurity surveillance: Quantitative approaches*, CABI Invasive series (pp. 167–180). Wallingford, Oxfordshire, UK: CABI.

Springborn, M. R., Lindsay, A. R., & Epanchin-Niell, R. S. (2016). Harnessing enforcement leverage at the border to minimize biological risk from international live species trade. *Journal of Economic Behavior & Organization*, *132*, 98–112.

Venette, R. C., Moon, R. D., & Hutchison, W. D. (2002). Strategies and statistics of sampling for rare individuals. *Annual Review of Entomology*, *47*(1), 143–174.

Wan, F.-H. & Yang, N.-W. (2016). Invasion and management of agricultural alien insects in China. *Annual Review of Entomology*, *61*(1), 77–98.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, *14*, 867–897.

Yamamura, K. & Sugimoto, T. (1995). Estimation of the pest prevention ability of the import plant quarantine in Japan. *Biometrics*, *51*(2), 482–490.