

1 GLM for partially pooled categorical predictors with a  
2 case study in biosecurity

3 Christopher M. Baker<sup>1,2,3</sup> Howard Bondell<sup>1,2</sup>

4 Nathaniel Bloomfield<sup>1,2,3</sup> Elena Tartaglia<sup>4</sup>

5 Andrew P. Robinson<sup>1,3</sup>

6 December 15, 2022

7 1. School of Mathematics and Statistics, The University of Melbourne

8 2. Melbourne Centre for Data Science, The University of Melbourne

9 3. Centre of Excellence for Biosecurity Risk Analysis, The University of Melbourne

10 4. Data61, Commonwealth Scientific and Industrial Research Organisation

11 cbaker1@unimelb.edu.au

12 howard.bondell@unimelb.edu.au

13 nathaniel.bloomfield@unimelb.edu.au

14 elena.tartaglia@data61.csiro.au

15 apro@unimelb.edu.au

16 **Acknowledgements**

17 We would like to thank Tom Waring for their contribution to editing this manuscript.

## Abstract

National governments use border information to efficiently manage the biosecurity risk presented by travel and commerce. In the Australian border biosecurity system, data about cargo entries are collected from records of directions: that is, the records of actions taken by the biosecurity regulator. An entry is a collection of import lines where each line is a single type of item or commodity. Analysis is simple when the data are recorded in line mode: the directions are recorded individually for each line. The challenge comes when data are recorded in container mode, because the same direction is recorded against each line in the entry, meaning that we don't know which line(s) within the entry are non-compliant. We develop a statistical model to use container mode data to help inform biosecurity risk of items. We use asymptotic analysis to estimate the value of container mode data compared to line mode data, do a simulation study to verify that we can accurately estimate parameters in a large dataset, and we apply our methods to a real dataset, for which important information about the risk of non-compliance is recovered using the new model.

*Keywords:* Bayesian inference; risk analysis; cargo

## 1 Introduction

Invasive species pose a multifaceted threat to society, leading to reductions in agricultural productivity, as well as damages to the environment, human health, and the economy [Kumar Rai and Singh, 2020]. Considerable effort is devoted to managing invasive species [Jardine and Sanchirico, 2018], in either eradicating them [Baker and Bode, 2020, Holmes et al., 2019, Wenger et al., 2017, Helmstedt et al., 2016], or suppressing their numbers to reduce damages [Binny et al., 2021, Brook et al., 2012, Sharov et al., 2002]. The costs associated with managing invasive species provides governments with an incentive to manage biosecurity risks at national borders to prevent the establishment of the new

44 species.

45       Given the massive scale of global trade, biosecurity regulators need to be able to  
46 allocate their resources efficiently, but to do so they must understand the risks associated  
47 with various entries. At the border, one of the most effective ways of achieving this is by  
48 using the outcome of previously conducted inspections as intelligence that informs future  
49 operations. This allows regulators to identify high risk commodities and importers, and  
50 modify their inspection targets and policies accordingly.

51       However, gaining insight from border inspections is a massive logistical challenge,  
52 as data must be consistently recorded in a format that makes this analysis possible.  
53 Putting infrastructure into place to collect this data, and collecting it accurately can  
54 be expensive and challenging. Often, shortcuts may be taken that render the data less  
55 valuable in analysing patterns of risk.

56       In Australia systems have been put in place to capture biosecurity data since the  
57 early 90's; these data are extracted from the directions applied to the cargo in question.  
58 Directions are used to control the movement and direct the assessment and management  
59 of goods subject to biosecurity control, and the *modes* discussed in this paper correspond  
60 to how those directions are carried out: at the line level, or at the entry level. As such,  
61 in line mode, the details of which items within an entry are inspected and found to  
62 be compliant or non-compliant are fully recorded. However, in container mode, the  
63 results of an inspection are applied to all items within an entry. Container mode was  
64 introduced so that containers could be released piecemeal in order to minimise the  
65 bottlenecks created in ports when lines that comprise many containers are held until  
66 all of the line has been cleared. Container mode makes it much quicker for border staff  
67 to manage entries with a large number of items, but means that when the data are  
68 analysed, entries in container mode are censored — in these cases, it is unknown which  
69 items were inspected, and which of those were found to be compliant or non-compliant.  
70 This makes analysis of the data to identify trends in biosecurity risk challenging.

71 Data that have been collected in container mode are closely related to the data  
 72 collected under *pooled testing*, which is often used for disease surveillance. Within the  
 73 pooled testing literature there are two main branches: one aims to identify positives  
 74 within a pool, while the other seeks to use pooled data to estimate quantities about  
 75 the population. It is the latter – estimating quantities – that we are interested in.  
 76 The fundamental problem is estimating a prevalence,  $p$ , within a population, when  
 77 only pooled data are available [Thompson, 1962]. More recent work has focused on  
 78 improving estimates by reducing bias, either through altering the sampling strategy  
 79 [Schaarschmidt, 2007, Hepworth and Watson, 2009] or by incorporating bias correction  
 80 into models [Hepworth and Biggerstaff, 2017, 2021]. There have also been extensions of  
 81 the problem where  $p$  is not a constant, but it is estimated using linear regression using  
 82 only the pooled data [Delaigle and Hall, 2015, Chatterjee and Bandyopadhyay, 2020,  
 83 McMahan et al., 2017, Liu et al., 2020]. These papers have made significant progress in  
 84 fitting increasingly complex models, but do not focus on the impacts of different types  
 85 of pooling on the precision of model estimates.

86 In practice, however, pooled testing is manifestly different from the biosecurity sce-  
 87 nario. Pooled testing exists by design: as a way to gather information about a population  
 88 while reducing testing. In biosecurity, inspection is applied to every individual line, and  
 89 the results are only pooled at the point of data capture — as a side effect of the mode  
 90 selection of the entry. Hence, we are interested in how much information we are losing  
 91 due to aggregating results as container mode. Our analysis offers regulators the oppor-  
 92 tunity to assess the risks of the continued use of container mode, and to weigh them  
 93 against its operational advantages.

94 In this paper we investigate the effect of container mode data collection upon our  
 95 ability to estimate the biosecurity risk of items. We start with an asymptotic analysis,  
 96 where we calculate the precision of estimates and determine the implications of mixing  
 97 different item types in container mode. We then develop a simulation experiment that

allows us to understand how larger entries and more item types affect the precision of our estimates. Finally, we analyse biosecurity data provided by the Australian Department of Agriculture, Fisheries and Forestry (DAFF) to identify the real-world differences between using the line-only data and including the container mode data.

## 2 Model overview

### 2.1 Data

To make our language about the data more precise, we will explicitly define what we mean by entries, lines and directions. Entries are a collection of lines, and a line is a group of the same type of item or commodity being imported. When cargo enters the country, each line is given *directions*. These directions detail all of the activities undertaken by the biosecurity regulator to manage the biosecurity risk of each line, and also the outcome of those activities. To isolate the uncertainty that arises from use of container mode, we focus on one aspect of the directions recorded against import lines: whether they were deemed compliant (within biosecurity regulations) or not.

Line and container mode are the two ways that records are kept of actions made by the biosecurity regulator. In line mode, the directions assigned to each line are recorded along with the outcome for that line. In container mode, directions are only recorded per entry. This means that in container mode, if any line in that entry has an inspection, and non-compliance is found, then every line in that entry is recorded to be non-compliant. If all of the lines in the entry are compliant, then they are all marked compliant, so in this case line and container mode are equivalent.

### 2.2 Modelling

Throughout this paper, we focus on estimating the probability that a line is non-compliant using information including the type of item, country of origin and whether

122 it has complete documentation. The full model for the probability that a line is non-  
 123 compliant,  $p_{ijk\ell}$ , is:

$$\text{logit}(p_{ijk\ell}) = \alpha_i + \beta \mathbb{I}_j + \delta_k + \gamma_\ell, \quad (1)$$

124 where the fixed effects are  $\alpha_i$ ,  $\beta \mathbb{I}_j$  and  $\delta_k$ :  $\alpha_i$  represents the item type, the indicator  
 125 variable  $\mathbb{I}_j$  denotes whether there is correct documentation, the coefficient  $\beta$  is the given  
 126 to the instance that there is correct documentation, and  $\delta_k$  represents the country of  
 127 provenance. The random effect  $\gamma_\ell$  represents the entry effect, which we include because  
 128 there may be correlations between lines within an entry. We anticipate some correlation  
 129 because lines in the same entry originate from a common context, so they may be non-  
 130 compliant for related reasons. The indices can take values

$$i = 1, \dots, a, \quad a \in \mathbb{N}, \quad a = \# \text{ items} \quad (2)$$

$$j = 1, 2, \quad \text{without and with documentation} \quad (3)$$

$$k = 1, \dots, d, \quad d \in \mathbb{N}, \quad d = \# \text{ countries} \quad (4)$$

$$\ell = 1, \dots, g, \quad g \in \mathbb{N}, \quad g = \# \text{ entries.} \quad (5)$$

131 The values of the indicator variable are

$$\mathbb{I}_1 = 0, \quad \text{without documentation} \quad (6)$$

$$\mathbb{I}_2 = 1, \quad \text{with documentation.} \quad (7)$$

132 The random effect  $\gamma_\ell$  has distribution

$$\gamma_\ell | \sigma \sim \text{Normal}(0, \sigma), \quad \ell = 1, \dots, g, \quad g \in \mathbb{N}. \quad (8)$$

133 If all data were in line mode then the above model would be a fairly standard mixed  
134 effects logistic regression with categorical variables. However, because of the use of  
135 container mode to capture the data, we don't observe outcomes for each line, as every  
136 line in the entry is marked as non-compliant if any line in the entry is found to be  
137 non-compliant. Therefore, the outcome is whether the entry is compliant and we need  
138 to calculate the probability that the entry is non-compliant, which is one minus the  
139 probability that every line in the entry is compliant:

$$\mathbb{P}(\text{Entry } l \text{ non-compliant}) = q_l = 1 - \prod_{ijk \text{ for lines in } l} [1 - p_{ijkl}], \quad (9)$$

140 where  $p_{ijkl}$  is the probability that the line with indices  $ijkl$  is non-compliant, calculated  
141 from Eq. (1). Hence, for entries in container mode, we treat the entry as a Bernoulli  
142 random variable with probability defined by Eq. (9), while for entries in line mode, we  
143 treat each line as a Bernoulli random variable with probability as defined in Eq. (1).

144 This paper includes three analyses: an asymptotic analysis, a simulation study, and a  
145 case study of Australian biosecurity data. For the asymptotic analysis we only consider  
146 the item type, ignoring effects due to the country of origin, documentation and entry  
147 effect. As such, rather than using Eq. (1), we just consider the probability that a line  
148 of item type  $i$  is non-compliant,  $p_i$ . The simulation study and the case study both use  
149 the full model, as defined above.

### 150 3 Asymptotic analysis

151 We use asymptotic analysis to investigate how the precision of estimates depends on  
152 entry size, the number of entries, the probability of non-compliance and whether item  
153 types are mixed. This analysis comprises two parts. The first assumes that all items  
154 are a single type, which allows us to quantify how the amount of data, probability of  
155 non-compliance and entry size affect precision. The second part assumes that there are

156 two different item types, and it explores how changing the proportion of entries with  
 157 both item types mixed affects precision.

158 Throughout this section we make two simplifications. Firstly, we do not separate  
 159 line mode and container mode because container mode data with an entry size of one  
 160 is mathematically equivalent to line mode data. Hence, throughout these analysis, an  
 161 entry size of one means line mode and entry size greater than one implies container  
 162 mode. Secondly, we assume that each item has a fixed probability of non-compliance.  
 163 As such, any uncertainty in the inference of this value arises as a result of the difference  
 164 between container and line mode. With this in hand, we consider each line a Bernoulli  
 165 trial, which only depends on the item type. When the entry size is greater than one,  
 166 the relevant probability is whether at least one line was non-compliant.

167 We estimate precision using an asymptotic estimate of the standard error. We cal-  
 168 culate the precision from the square roots of the diagonal elements in the Fisher infor-  
 169 mation matrix,  $\mathcal{I}$ , which is the expected value of the negative of the Hessian matrix of  
 170 the log-likelihood evaluated at the value of the parameter.

### 171 3.1 Single item type

172 For the single item type case, we set the probability of non-compliance to be  $p$ , and define  
 173  $N$  as the total number of entries,  $I$  as the number of non-compliant entries and  $S$  as the  
 174 size (i.e. the number of lines) in each entry. The likelihood is a binomial distribution,  
 175 where the outcome is the discovery of a non-compliant entry. The probability that an  
 176 entry is compliant is

$$\mathbb{P}(\text{entry compliant}) = (1 - p)^S, \quad (10)$$

177 meaning that the binomial likelihood for a set entry size  $S$  proportional to



$$\mathcal{L}_S = (1 - (1 - p)^S)^I (1 - p)^{S(N-I)}. \quad (11)$$

178 Therefore, the log-likelihood is

$$\log \mathcal{L}_S = I \log (1 - (1 - p)^S) + (N - I) \log ((1 - p)^S). \quad (12)$$

179 As there is only one parameter, we calculate its second derivative (rather than needing  
180 a Hessian matrix):

$$\left[ \frac{\partial^2 \log \mathcal{L}_S}{\partial p^2} \right] = \frac{S \left( N + \frac{I((1+S)(1-p)^{S-1})}{((1-p)^S - 1)^2} \right)}{(1 - p)^2}. \quad (13)$$

181 To calculate the Fisher information, we take the expectation of the number of entries  
182 with non-compliance, which depends on the size of the entry:

$$\mathbb{E}[I] = N(1 - (1 - p)^S). \quad (14)$$

183 Hence the Fisher information is

$$\mathcal{I} = -\mathbb{E} \left[ \frac{\partial^2 \log \mathcal{L}_S}{\partial p^2} \right] = -\frac{NS^2(1 - p)^{S-2}}{(1 - p)^S - 1}, \quad (15)$$

184 and the standard error estimate is

$$SE = \left( -\frac{NS^2(1 - p)^{S-2}}{(1 - p)^S - 1} \right)^{-1/2}. \quad (16)$$

185 Using Eq. (16) we can understand how the probability of non-compliance, the entry  
186 size and the number of lines affect the standard error, and we plot these relationships  
187 in Figure 1. The left plot shows that the standard error depends on the probability

188 of non-compliance and that the relationship depends on the entry size. For all entry  
 189 sizes, the standard error is small when the probability of non-compliance is small (below  
 190  $\sim 0.3$ ). However, for larger values of the probability of non-compliance, the standard  
 191 error increase significantly if the entry size is three or greater. The large  $p$  behaviour is  
 192 driven by the  $(1 - p)^{S-2}$  term in Eq. (16), which means SE goes to 0 if  $S = 1$ , while it  
 193 diverges if  $S \geq 3$ . Figure 1 also shows how the standard error decreases as the number  
 194 of lines of data increases. The lower the entry size is, the lower the standard error, and,  
 195 as  $SE \sim \sqrt{1/N_{E,S}}$ , container mode data with larger entry sizes require a large amount  
 196 of data to reach the same standard error.

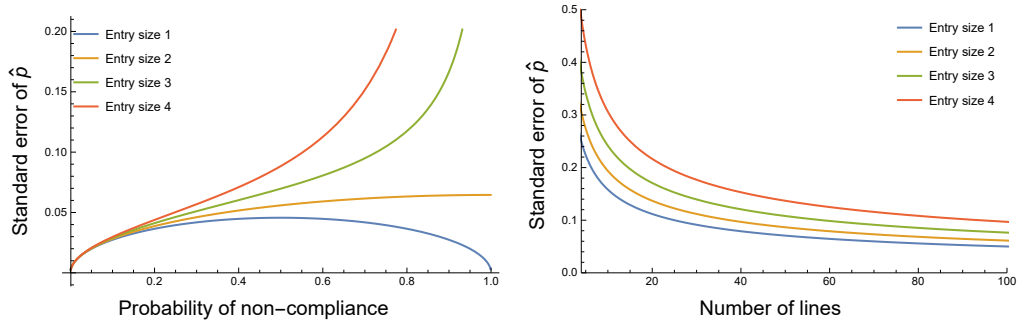


Figure 1: The standard error (Eq. (16)) as the probability of non-compliance,  $p$ , is varied  
 (left) and as the number of lines are varied (right). For the left plot the number of lines  
 is held constant at 120. For the right plot, the probability of non-compliance is held  
 constant at 0.5

### 197 **3.2 Two item types**

198 In this section we consider a situation where there are two items with probabilities of  
 199 non-compliance of  $p_1$  and  $p_2$ , and we examine how these different items probabilities  
 200 interact. We focus on a scenario where every entry is of size two, meaning there are  
 201 three types of entries: only type 1; only type 2; or mixed, with one line of type 1 and  
 202 one of type 2. We denote the number of lines within a single entry of type 1 and 2

as  $S_1$  and  $S_2$  respectively, and  $I(S_1, S_2)$  and  $N(S_1, S_2)$  are the number of entries with non-compliance and total number of entries with  $S_1$  type 1 lines and  $S_2$  type 2 lines. For our case, we can have  $S_1 = 2, S_2 = 0$ ;  $S_1 = 1, S_2 = 1$ ; or  $S_1 = 0, S_2 = 2$ . Rewriting the log-likelihood from Eq. (12), we get

$$\log \mathcal{L} = \sum_{S_1, S_2} I(S_1, S_2) \log \left( 1 - (1 - p_1)^{S_1} (1 - p_2)^{S_2} \right) + (N(S_1, S_2) - I(S_1, S_2)) \log \left( (1 - p_1)^{S_1} (1 - p_2)^{S_2} \right). \quad (17)$$

We compute the Fisher information matrix and the standard error using Mathematica. The standard error for  $p_1$  is

$$\frac{1}{2} \sqrt{-\frac{p_1 (p_1^2 - 3p_1 + 2) (N_{1,2}(p_1 - 1)(p_2 - 2)p_2 + 4N_{2,2}(p_2 - 1)(p_1(p_2 - 1) - p_2))}{N_{1,1}(p_1 - 1)(N_{1,2}(p_1 - 1)(p_2 - 2)p_2 + 4N_{2,2}(p_2 - 1)(p_1(p_2 - 1) - p_2)) + N_{1,2}N_{2,2}(p_1 - 2)p_1(p_2 - 1)^2}}, \quad (18)$$

where  $N_{1,1}$  and  $N_{2,2}$  are the number of entries with two type 1 lines and two type 2 lines respectively, while  $N_{1,2}$  are the number of entries with both type 1 and type 2 lines. The standard error for  $p_2$  is the same, with  $N_{1,1}$  and  $N_{2,2}$  switched and  $p_1$  and  $p_2$  switched. By examining Eq. (18) we can see that the behaviour of the standard errors is more complex than when we considered only one type of item. Notably, the number of entries of only type 2,  $N_{2,2}$ , is in the equation for the type 1 standard error, along with the probability of non-compliance of type 2,  $p_2$ .

Figure 2 shows how the standard error varies with the proportion of mixed entries, for different non-compliance probabilities. Here we fix  $N_{\text{total}} = 50$ , and keep  $N_{1,1} = N_{2,2}$  while the proportion  $N_{1,2}/N_{\text{total}}$  is varied. In each case, we hold  $p_1$  at 0.1, and we vary  $p_2$  for 0.05 up to 0.9 across the four plots. The way that the standard error changes as a function of the proportion of mixed entries changes markedly, depending on the value of  $p_2$ . In particular, when  $p_2$  is 0.7 and 0.9, the standard error for  $p_1$  increases as the proportion of mixed entries increases, while the standard error for  $p_2$  actually decreases,

223 while the proportion of mixed entries is below  $\sim 90\%$ .

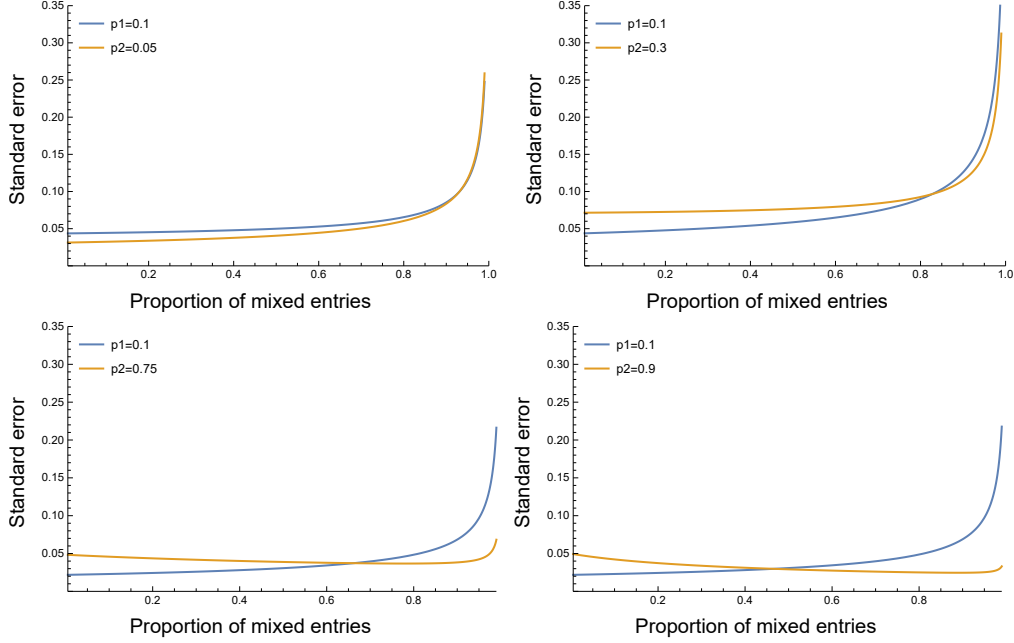


Figure 2: The standard error when for entry size 2, when the proportion of mixed entries is varied. In each plot there are 50 total entries and  $p_1$  is held constant at 0.1, while  $p_2$  goes from 0.05 up to 0.9.

### 224 3.3 Asymptotic analysis conclusions

225 From the asymptotic analysis we make two observations:

- 226 1. Container mode data are most useful when the probability of non-compliance,  $p$ ,  
227 is small. In the extreme case where  $p \rightarrow 0$ , container mode data are equivalent  
228 to line mode data because whenever an entry is found to be compliant, we know  
229 that every line within that entry is compliant, regardless of whether it is container  
230 mode or line mode.
- 231 2. By allowing entries to contain lines of different item types, container mode can  
232 increase or decrease the precision of estimates, depending on the true probability  
233 of non-compliance.

## 234 4 Simulation study

235 As the first step in developing a method to analyse real-world data sets, we simulate a  
236 data set that contains key complexities including multiple item types, fixed effects and  
237 random effects. Once we have simulated data, we can fit a model to estimate parameters.  
238 The advantage of our approach is that we can (1) verify that our model behaves correctly  
239 and (2) explore how model precision varies in a more complex setting.

### 240 4.1 Data simulation

241 We simulate data from the full model (as described in Section 2.2). For our simulations,  
242 we chose parameters that we expect to be similar to real-world values. We set  $\beta = -1$ ,  
243 choosing an negative effect of having correct documentation, because we expect lines  
244 with correct documentation have a higher chance of being compliant. We include  $a = 5$   
245 item types with  $\alpha_i$  taking values of -6.91, -4.60, -3.89, -2.94, -1.39 (corresponding to  
246 non-compliance probabilities of 0.001, 0.01, 0.02, 0.05 and 0.2, if all other effects were  
247 0). We chose negative values for the item effect  $\alpha_i$ , since the probability of detection  
248 is expected to be low for any item in real-world data. We used  $d = 3$  countries and  
249 set their weights to be 0.5, -1 and 0.25. For within-entry correlation, we set  $\sigma = 0.25$ .  
250 For each entry we draw whether it is in line mode or not with probability 0.25, and for  
251 each line we set the probability of having correct documentation to be 0.2. We show an  
252 example of the format of the data in Table 1.

### 253 4.2 STAN model

254 We model the system in a Bayesian framework, using the RSTAN package in R. We  
255 choose a Bayesian package due to the ease of specifying the model and could be updated,  
256 in principle, for any specification of the model for the probability of non-compliance.  
257 The basic structure for fitting the simulated data is to define the probability of non-

Table 1: Example simulated data. Lines 1-3 are all marked as non-compliant because entry 1 is in container mode, even though only 1 or 2 of the lines were actually non-compliant.

Line	Entry	Type	Documentation	Mode	Non-compliant
1	1	3	1	Container	1
2	1	2	0	Container	1
3	1	1	0	Container	1
4	2	5	0	Line	0
5	3	4	1	Line	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

258 compliance for each line (following Eq. (1))

$$\text{logit}(p_{ijk\ell}) = \alpha_i + \beta \mathbb{I}_j + \delta_k + \gamma_\ell, \quad (19)$$

259 where

$$\gamma_\ell | \sigma \sim \text{Normal}(0, \sigma), \quad \ell = 1, \dots, g, \quad g \in \mathbb{N}. \quad (20)$$

260 As we fit our model in a Bayesian framework, we set priors on each parameter:

$$\alpha_i \sim \text{Normal}(-4, 4), \quad i = 1, \dots, a, \quad a \in \mathbb{N}, \quad (21)$$

$$\beta \sim \text{Normal}(0, 0.5), \quad (22)$$

$$\delta_k \sim \text{Normal}(0, 0.5), \quad k = 1, \dots, d, \quad d \in \mathbb{N} \quad (23)$$

$$\sigma \sim \text{Normal}(0, 0.5) \quad (24)$$

261 The probability  $p_{ijkl}$  is the probability of non-compliance of a line that has charac-  
 262 teristics  $(i, j, k, \ell)$ . We denote the probability that line  $n$  is non-compliant as  $p_{(n)}$ , where  
 263  $p_{(n)} = p_{ijkl}$  if that line has characteristics  $(i, j, k, \ell)$ .

264 We fit the STAN model to the simulated data for two reasons. Firstly we confirm  
 265 that, with sufficient data, the STAN model can accurately estimate model parameters.  
 266 Secondly we explore how changing the amount of data and the proportion of container  
 267 mode data affects precision, over a range of entry sizes. To measure precision, we use  
 268 the standard deviation of the posterior samples, because it is the Bayesian equivalent of  
 269 the standard error calculations in Section 3.

### 270 4.3 Simulation results

271 Figure 3 shows that the STAN model gives good parameter estimates, if there is  
 272 sufficient data. Overall, the STAN estimates are close to the true values when there is a  
 273 large amount of data. The two smallest values of  $\alpha$  show the worst performance, which  
 274 is not surprising, given our asymptotic analysis results (Section 3).

275 The precision of model estimates depends on the entry size, the probability of non-  
 276 compliance and the amount of container mode data (Figure 4). The best-case scenario  
 277 is when all data are in line mode (ratio of container mode = 0), and the distance from  
 278 that line shows the impact of using a mix, or only, container mode data. We find that  
 279 the difference between the precision estimates depends strongly on the combination of  
 280 factors. There are some combinations where all analyses return similar precision (e.g.  
 281  $\alpha = -1.4$  and entry size 10), while other combinations have a large gap where the line  
 282 only data far outperforms the others (e.g.  $\alpha = -4.6$  and entry size = 5).

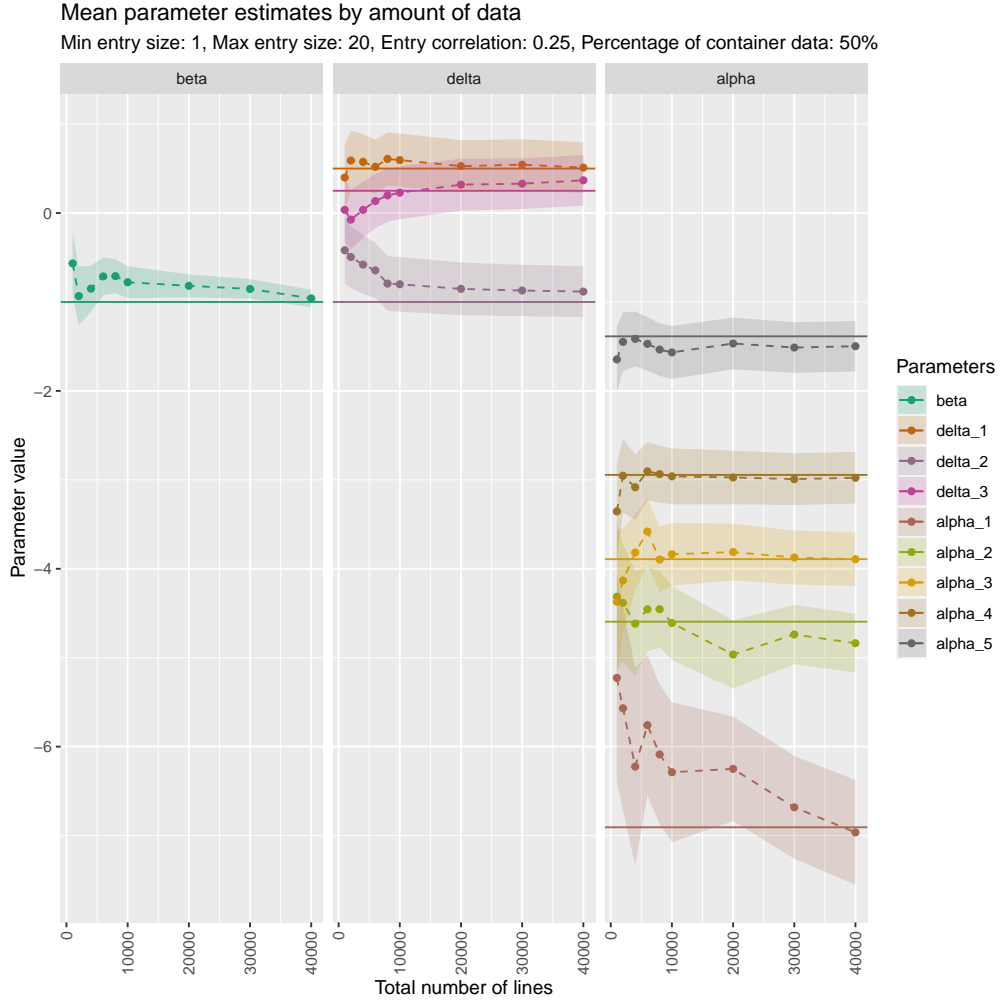


Figure 3: Results from the STAN model, fitted to simulated data. 50% of the data are in container mode and 50% are in line mode, and each entry is a random size between 1 and 20. Dots and dashed lines in each plot show the STAN estimates, while the solid lines show the true value of each parameter.

## 283 5 Case study

284 We apply our model to some real biosecurity data, to see how our understanding of  
285 the system changes by fitting the full data, compared to only using the container mode  
286 data. The data set is of furniture imports in 2020, and we break the data set into 52  
287 weekly data sets for this analysis. Because container mode data with no non-compliance



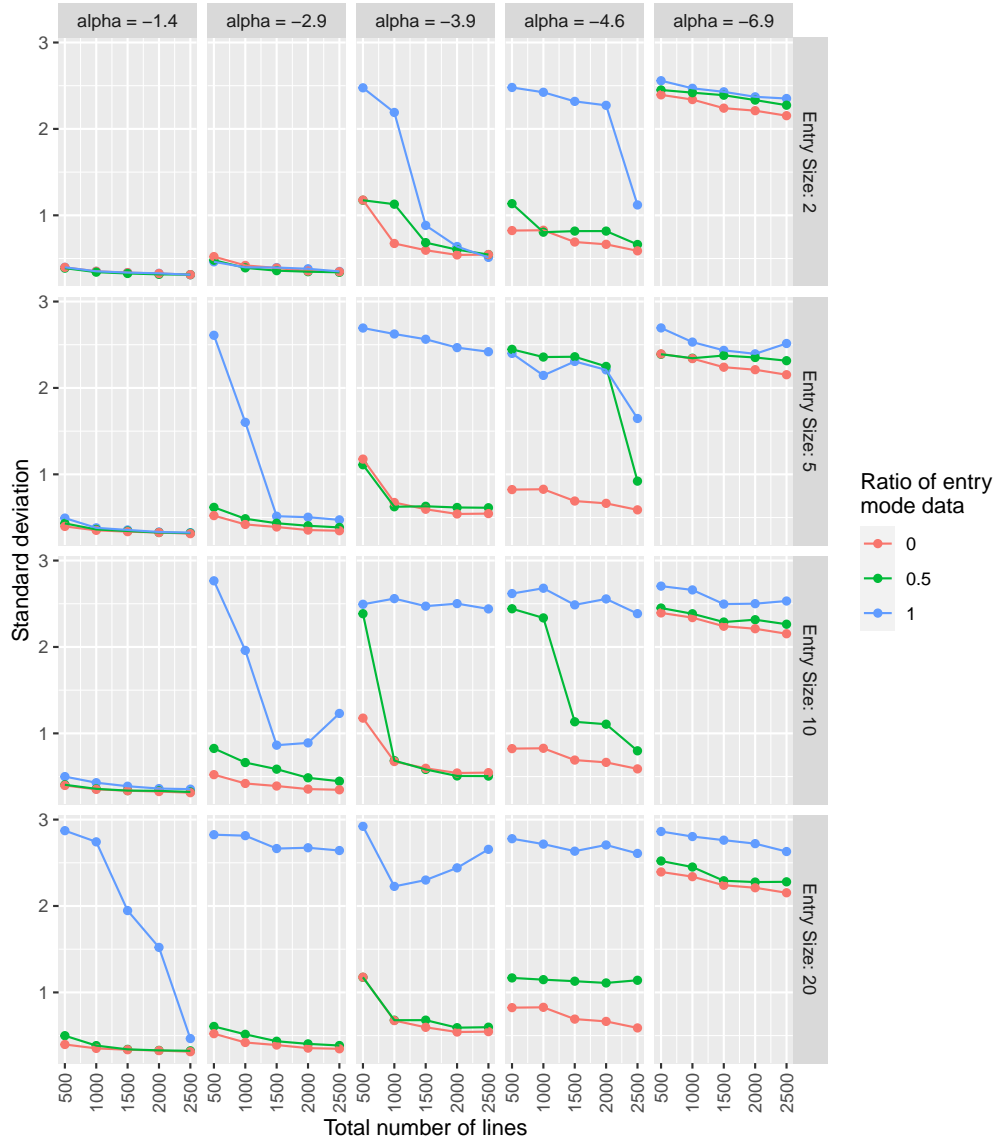


Figure 4: The standard deviation of the alpha parameter estimates for simulations with varying entry size and total number of lines. The ratio of container mode data is the proportion of data in container mode, meaning that 0 corresponds to line mode data-sets, 0.5 if a mixed data-set and 1 is purely container mode.

is equivalent to line mode data, the more container-mode entries with non-compliance,  
the larger the differences could be between analysing all of the data compared to only  
the line mode data. There is a natural break in the dataset, where three weeks have

291 four container mode entries with non-compliance, with the remaining weeks having fewer.  
 292 Hence, we choose to analyse these three weeks as a demonstration of potential real-world  
 293 differences between analysing all data and only line mode data. For confidentiality, for  
 294 each week we re-name the item type, the country and the week to be integers (week 1,  
 295 2 and 3 do not correspond to the first 3 weeks of the year, and country 1 and item 1 are  
 296 not the same in weeks 1 and 2). Relabeling the data does not cause issues because we  
 297 are not trying to compare estimates between weeks or draw inferences between countries  
 298 or item types in this analysis. A summary of the data is given in Table 2.

Table 2: The number of lines in container mode and line mode for each week. The ratio  
 is the fraction of the week’s data that is in container mode.

Week	Entry	Line	Ratio
1	516	6741	0.07
2	1667	7213	0.19
3	679	6832	0.09

299 We fit our STAN model to the data from each week, using all the data and the  
 300 line-only subset of the data. Because the line-only analysis is a subset of the data, not  
 301 every country or item is present in the line-only data set. For the parameters that do  
 302 match, we generate a scatter plot of the mean estimates to see how frequently we get  
 303 different estimates (Figure 5). For many of the parameters, we get very similar results  
 304 irrespective of which data we use. This is not surprising, given that most of the data  
 305 are in line mode and that line mode data gives more information than container mode  
 306 data. However, it is interesting that including the container mode data results in quite  
 307 large changes to some of the parameter estimates.

308 Analysing the full data set, rather than the line-only data, gives us information about  
 309 more items and countries. Table 3 shows all of the parameter estimates that we only

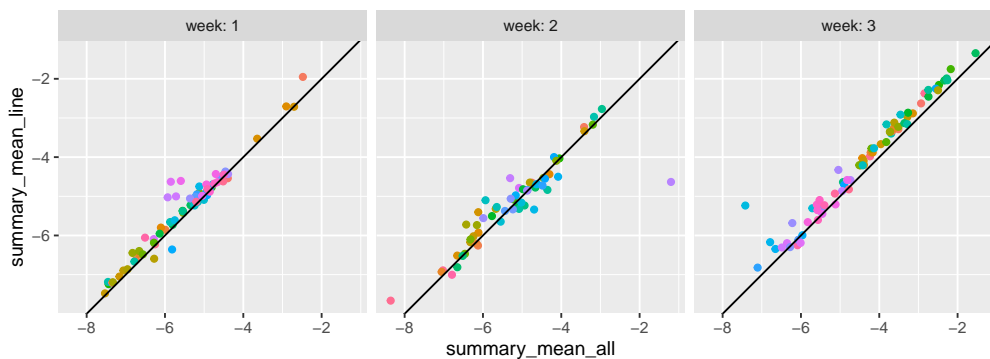


Figure 5: Scatter plot of parameter estimates using the line only data and all data. Dots that fall on the solid black line indicate when we get the same parameter estimate from each data set.

get when including the container mode data in the analysis.

## 6 Discussion

In this paper we sought to understand how container mode data affects our ability to estimate risks in the biosecurity setting. Due to the relative ease of analysing line data, in our experience, container mode data are often excluded from data analysis. From our analysis of real biosecurity data, we find that including the container mode data in analysis can markedly change some results and that by only using line data the analysis is somewhat limited – there are parameter estimates that simply cannot be made.

Through our simulation study and our asymptotic analysis we gain an understanding of how mixing container mode data into line mode data impacts our ability to make precise inferences. We see that mixing lines that have different probabilities of being non-

Table 3: Parameter estimates that are only possible when using all data, as opposed to the line-only fits. The value `summary_rhat_all` measures convergence of MCMC chains [Brooks et al., 2011].

	param_name	week	summary_mean_all	summary_sd_all	summary_rhat_all
1	p_intercept[72]	1.00	-1.70	1.68	1.00
2	p_intercept[59]	2.00	-1.34	1.71	1.01
3	p_intercept[60]	2.00	-2.07	1.63	1.04
4	p_intercept[61]	2.00	-1.18	1.92	1.00
5	country_effect[62]	2.00	-0.01	0.49	1.00
6	country_effect[63]	2.00	-0.01	0.47	1.00
7	country_effect[64]	2.00	-0.01	0.49	1.01
8	country_effect[65]	2.00	0.02	0.48	1.02
9	country_effect[66]	2.00	0.01	0.48	1.00
10	country_effect[67]	2.00	-0.01	0.48	1.00
11	p_intercept[72]	3.00	-1.43	1.78	1.00

compliant has divergent affects on model precision. In the asymptotic analysis we see that mixing a high-probability item with a lower probability item makes it easier to precisely estimate the high-probability item’s parameter, while mixing two low-probability items makes it challenging to make inferences about either. When one item has a low probability of non-compliance, a non-compliant mixed entry is likely to correspond to a non-compliant entry of the of the item with a higher probability of compliance. The extra data from this entry can therefore be used to refine the parameter estimates, whereas if the respective probabilities are not well resolved this is not possible. These interactions between items are exacerbated in the simulation study, where we have larger entries and more item types, and we see differing patterns of item types, entry sizes and amounts

331 of data where the presence of container mode data degrades model precision, compared  
332 to a line-only analysis.

333 While we are able to model the full dataset, the continuing presence of container  
334 mode data will be a barrier for future data analysis. Most statistical and machine learn-  
335 ing algorithms are designed such that each row of data will have explanatory variables  
336 and an outcome. With container mode data, we have the explanatory variables, but only  
337 partial information about the outcome. Hence, we need customised algorithms (such as  
338 described in this paper) to analyse it appropriately. Even so, many simplifying assump-  
339 tions have been made: for example, that a given item’s probability of non-compliance  
340 does not vary over time. These simplifications do not change our conclusions, as the  
341 analysis still serves to highlight the differences between container and line mode. How-  
342 ever, without proper care, the specific results of our case-study analysis should not be  
343 applied to decision making.

344 While in our case the pooling of outcomes in container mode is an artifact of the  
345 system, pooled testing are often designed strategically in public health to improve effi-  
346 ciency. The strain on the PCR testing system has prompted regimes to identify cases  
347 while minimising the rounds of testing [Mutesa et al., 2021]. While it is clear that the  
348 way pooling is conducted affects how quickly infections can be identified, in some cir-  
349 cumstances, the outcomes of pooled testing can also be used to estimate parameters,  
350 and our work has implications for these situations [Delaigle and Hall, 2015, Chatterjee  
351 and Bandyopadhyay, 2020, McMahan et al., 2017, Liu et al., 2020]. From our asymptotic  
352 analysis, it is clear that even when only estimating a prevalence, for a fixed number of  
353 tests, the standard error of the estimate depends on pool size and therefore there would  
354 be an optimal pool size, which would depend on the prevalence [Keeling and Rohani,  
355 2008]. When there is a model with parameters being estimated, the added complexity  
356 could compound the issue, making pool sizing a more important aspect of study design.  
357 Furthermore, as we demonstrate here, it is not only pool size that matters, it is how

358 samples with different characteristics are grouped together. Hence, for a given study,  
359 there may not be one optimal pooling strategy, and there would likely be trade-offs  
360 between different parameters when it comes to maximising the precision of estimates.

361 In this paper we have investigated how pooled data can reduce our ability to make  
362 statistical inferences about the population. Within the Australian biosecurity context,  
363 the pooling is due to how data are recorded and was implemented for operational reasons.  
364 Due to this operational decision, analysing full datasets is harder than the line-only mode  
365 data, meaning that either analysis is restricted to a subset of possible methods or that  
366 container mode data is ignored. While it does not follow that container mode should  
367 be removed due to operational efficiency, if the line-level data could be captured then it  
368 would improve our understanding of biosecurity risk. However, pooled data will continue  
369 to be collected in various fields, for example due to the cost-savings of testing multiple  
370 samples at once for diseases. While there is extensive work in pooled testing protocols  
371 for case identification, there is less work on identifying pool sizes when aiming to make  
372 inferences about aspects of the population. Our work shows that pooling tests with  
373 variable underlying prevalence affects precision differentially. Hence, careful considering  
374 should be given to designing pools when the data will be used to make inferences about  
375 the population.

## 376 7 Code availability

377 All code is available at <https://github.com/cmbaker00/container-line-analysis-public>.

## 378 8 Data availability

379 The data that support the findings of this simulation study are openly available at  
380 <https://github.com/cmbaker00/container-line-analysis-public>. The case study

381 data are available from the Australian Department of Agriculture, Fisheries and Forestry.  
382 Restrictions apply to the availability of these data, which were used under license for  
383 this study.

## 384 9 Conflict of interest

385 The authors declare no conflict of interest.

## 386 References

- 387 C. M. Baker and M. Bode. Recent advances of quantitative modeling to sup-  
388 port invasive species eradication on islands. *Conservation Science and Prac-*  
389 *tice*, n/a(n/a):e246, 2020. ISSN 2578-4854. doi: 10.1111/csp2.246. URL  
390 <https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/csp2.246>. \_eprint:  
391 <https://conbio.onlinelibrary.wiley.com/doi/pdf/10.1111/csp2.246>.
- 392 R. N. Binny, J. Innes, N. Fitzgerald, R. Pech, A. James, R. Price, C. Gillies,  
393 and A. E. Byrom. Long-term biodiversity trajectories for pest-managed ecolog-  
394 ical restorations: eradication vs. suppression. *Ecological Monographs*, n/a(n/a),  
395 2021. ISSN 1557-7015. doi: <https://doi.org/10.1002/ecm.1439>. URL <https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecm.1439>. \_eprint:  
396 <https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecm.1439>.
- 398 L. A. Brook, C. N. Johnson, and E. G. Ritchie. Effects of predator control on be-  
399 haviour of an apex predator and indirect consequences for mesopredator suppres-  
400 sion. *Journal of Applied Ecology*, 49(6):1278–1286, Dec. 2012. ISSN 1365-2664.  
401 doi: 10.1111/j.1365-2664.2012.02207.x. URL [http://onlinelibrary.wiley.com.](http://onlinelibrary.wiley.com/ezp.lib.unimelb.edu.au/doi/10.1111/j.1365-2664.2012.02207.x/abstract)  
402 [ezp.lib.unimelb.edu.au/doi/10.1111/j.1365-2664.2012.02207.x/abstract](http://onlinelibrary.wiley.com/ezp.lib.unimelb.edu.au/doi/10.1111/j.1365-2664.2012.02207.x/abstract).

- 403 S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte*  
404 *Carlo*. CRC press, 2011.
- 405 A. Chatterjee and T. Bandyopadhyay. Regression models for group testing: Identi-  
406 fiability and asymptotics. *Journal of Statistical Planning and Inference*, 204:141–  
407 152, Jan. 2020. ISSN 0378-3758. doi: 10.1016/j.jspi.2019.05.003. URL <https://www.sciencedirect.com/science/article/pii/S0378375818301423>.
- 409 A. Delaigle and P. Hall. Nonparametric methods for group testing data, taking dilution  
410 into account. *Biometrika*, 102(4):871–887, Dec. 2015. ISSN 0006-3444. doi: 10.1093/  
411 biomet/asv049. URL <https://doi.org/10.1093/biomet/asv049>.
- 412 K. J. Helmstedt, J. D. Shaw, M. Bode, A. Terauds, K. Springer, S. A. Robinson, and  
413 H. P. Possingham. Prioritizing eradication actions on islands: it’s not all or nothing.  
414 *Journal of Applied Ecology*, 53:733–741, Jan. 2016. ISSN 1365-2664. doi: 10.1111/  
415 1365-2664.12599. URL [http://onlinelibrary.wiley.com.ezp.lib.unimelb.edu.](http://onlinelibrary.wiley.com.ezp.lib.unimelb.edu.au/doi/10.1111/1365-2664.12599/abstract)  
416 [au/doi/10.1111/1365-2664.12599/abstract](http://onlinelibrary.wiley.com.ezp.lib.unimelb.edu.au/doi/10.1111/1365-2664.12599/abstract).
- 417 G. Hepworth and B. J. Biggerstaff. Bias Correction in Estimating Proportions by Pooled  
418 Testing. *Journal of Agricultural, Biological and Environmental Statistics*, 22(4):602–  
419 614, Dec. 2017. ISSN 1537-2693. doi: 10.1007/s13253-017-0297-2. URL <https://doi.org/10.1007/s13253-017-0297-2>.
- 421 G. Hepworth and B. J. Biggerstaff. Bias Correction in Estimating Proportions by Imper-  
422 fect Pooled Testing. *Journal of Agricultural, Biological and Environmental Statistics*,  
423 26(1):90–104, Mar. 2021. ISSN 1537-2693. doi: 10.1007/s13253-020-00411-5. URL  
424 <https://doi.org/10.1007/s13253-020-00411-5>.
- 425 G. Hepworth and R. Watson. Debaised estimation of proportions in group test-  
426 ing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*,  
427 58(1):105–121, 2009. ISSN 1467-9876. doi: 10.1111/j.1467-9876.2008.00639.



x. URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9876.2008.00639.x>. eprint: <https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9876.2008.00639.x>.

N. D. Holmes, D. R. Spatz, S. Oppel, B. Tershy, D. A. Croll, B. Keitt, P. Genovesi, I. J. Burfield, D. J. Will, A. L. Bond, A. Wegmann, A. Aguirre-Muñoz, A. F. Raine, C. R. Knapp, C.-H. Hung, D. Wingate, E. Hagen, F. Méndez-Sánchez, G. Rocamora, H.-W. Yuan, J. Fric, J. Millett, J. Russell, J. Liske-Clark, E. Vidal, H. Jourdan, K. Campbell, K. Springer, K. Swinnerton, L. Gibbons-Decherong, O. Langrand, M. d. L. Brooke, M. McMinn, N. Bunbury, N. Oliveira, P. Sposimo, P. Geraldès, P. McClelland, P. Hodum, P. G. Ryan, R. Borroto-Páez, R. Pierce, R. Griffiths, R. N. Fisher, R. Wanless, S. A. Pasachnik, S. Cranwell, T. Micol, and S. H. M. Butchart. Globally important islands where eradicating invasive mammals will benefit highly threatened vertebrates. *PLOS ONE*, 14(3):e0212128, Mar. 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0212128. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0212128>.

S. L. Jardine and J. N. Sanchirico. Estimating the cost of invasive species control. *Journal of Environmental Economics and Management*, 87:242–257, Jan. 2018. ISSN 0095-0696. doi: 10.1016/j.jeem.2017.07.004. URL <http://www.sciencedirect.com/science/article/pii/S0095069616301322>.

M. J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008. ISBN 9780691116174. URL <http://www.jstor.org/stable/j.ctvcn4gk0>.

P. Kumar Rai and J. S. Singh. Invasive alien plant species: Their impact on environment, ecosystem services and human health. *Ecological Indicators*, 111:106020, Apr. 2020. ISSN 1470-160X. doi: 10.1016/j.ecolind.2019.106020. URL <https://www.sciencedirect.com/science/article/pii/S1470160X19310167>.

454 Y. Liu, C. S. McMahan, J. M. Tebbs, C. M. Gallagher, and C. R. Bilder. Generalized  
 455 additive regression for group testing data. *Biostatistics*, (kxaa003), Feb. 2020. ISSN  
 456 1465-4644. doi: 10.1093/biostatistics/kxaa003. URL [https://doi.org/10.1093/](https://doi.org/10.1093/biostatistics/kxaa003)  
 457 [biostatistics/kxaa003](https://doi.org/10.1093/biostatistics/kxaa003).

458 C. S. McMahan, J. M. Tebbs, T. E. Hanson, and C. R. Bilder. Bayesian regression  
 459 for group testing data. *Biometrics*, 73(4):1443–1452, 2017. ISSN 1541-0420. doi:  
 460 10.1111/biom.12704. URL [https://onlinelibrary.wiley.com/doi/abs/10.1111/](https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12704)  
 461 [biom.12704](https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12704). eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12704>.

462 L. Mutesa, P. Ndishimye, Y. Butera, J. Souopgui, A. Uwineza, R. Rutayisire, E. L.  
 463 Ndoricimpaye, E. Musoni, N. Rujeni, T. Nyatanyi, E. Ntagwabira, M. Semakula,  
 464 C. Musanabaganwa, D. Nyamwasa, M. Ndashimye, E. Ujeneza, I. E. Mwikarago,  
 465 C. M. Muvunyi, J. B. Mazarati, S. Nsanzimana, N. Turok, and W. Ndifon. A pooled  
 466 testing strategy for identifying SARS-CoV-2 at low prevalence. *Nature*, 589(7841):276–  
 467 280, Jan. 2021. ISSN 1476-4687. doi: 10.1038/s41586-020-2885-5. URL [https://](https://www.nature.com/articles/s41586-020-2885-5)  
 468 [www.nature.com/articles/s41586-020-2885-5](https://www.nature.com/articles/s41586-020-2885-5). Number: 7841 Publisher: Nature  
 469 Publishing Group.

470 F. Schaarschmidt. Experimental design for one-sided confidence intervals or hypothesis  
 471 tests in binomial group testing. *Communications in Biometry and Crop Science*, 2007.  
 472 ISSN 1896-0782. URL [http://agrobiol.sggw.waw.pl/~cbcs/pobierz.php?plik=](http://agrobiol.sggw.waw.pl/~cbcs/pobierz.php?plik=CBCS_2_1_5.pdf)  
 473 [CBCS\\_2\\_1\\_5.pdf](http://agrobiol.sggw.waw.pl/~cbcs/pobierz.php?plik=CBCS_2_1_5.pdf). Publisher: Faculty of Agriculture and Biology, Warsaw Agricultural  
 474 University, Poland.

475 A. Sharov, D. Leonard, A. Liebhold, E. Roberts, and W. Dickerson. “Slow The Spread”:  
 476 A National Program to Contain the Gypsy Moth. *Journal of Forestry*, 100(5):30–36,  
 477 July 2002.

478 K. H. Thompson. Estimation of the Proportion of Vectors in a Natural Population of

479 Insects. *Biometrics*, 18(4):568–578, 1962. ISSN 0006-341X. doi: 10.2307/2527902.  
480 URL <https://www.jstor.org/stable/2527902>. Publisher: [Wiley, International  
481 Biometric Society].

482 A. S. Wenger, V. M. Adams, G. D. Iacona, C. Lohr, R. L. Pressey, K. Morris,  
483 and I. D. Craigie. Estimating realistic costs for strategic management planning  
484 of invasive species eradication on islands. *Biological Invasions*, pages 1–19, Nov.  
485 2017. ISSN 1387-3547, 1573-1464. doi: 10.1007/s10530-017-1627-6. URL [https:](https://link.springer.com/article/10.1007/s10530-017-1627-6)  
486 [//link.springer.com/article/10.1007/s10530-017-1627-6](https://link.springer.com/article/10.1007/s10530-017-1627-6).