Submitted to the Annals of Applied Statistics

GLM FOR PARTIALLY POOLED CATEGORICAL PREDICTORS WITH A CASE STUDY IN BIOSECURITY

BY CHRISTOPHER M. BAKER[1,2,a], HOWARD BONDELL[2,d], EVELYN MANNIX[1,2,b], ELENA TARTAGLIA[3,f], THOMAS K. WARING[2,e] AND ANDREW P. ROBINSON[1,c]

[1]*Centre of Excellence for Biosecurity Risk Analysis, The University of Melbourne,* [a]*cbaker1@unimelb.edu.au;* [b]*evelyn.mannix@unimelb.edu.au;* [c]*apro@unimelb.edu.au*

[2]*Melbourne Centre for Data Science, The University of Melbourne,* [d]*howard.bondell@unimelb.edu.au;* [e]*tom.waring@unimelb.edu.au*

[3]*Data61, Commonwealth Scientific and Industrial Research Organisation,* [f]*elena.tartaglia@data61.csiro.au*

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.

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, Johnson and Ritchie (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 species.

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

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

Keywords and phrases: Bayesian inference, risk analysis, cargo.

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

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

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

In this paper we investigate the effect of container mode data collection upon our ability to estimate the biosecurity risk of items. We start with an asymptotic analysis, where we calculate the precision of estimates and determine the implications of mixing 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 it has complete documentation. In doing so, we make the assumption that the data are accurate – i.e. we don't consider imperfect detection. Therefore, The full model for the probability that a line is non-compliant, $p_{ijk\ell}$, is:

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

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

$$\begin{aligned} &i = 1, \ldots, a, \quad a \in \mathbb{N}, && a = \text{\# items} \\ &j = 1, 2, && \text{without and with documentation} \\ &k = 1, \ldots, d, \quad d \in \mathbb{N}, && d = \text{\# countries} \\ &\ell = 1, \ldots, g, \quad g \in \mathbb{N}, && g = \text{\# entries.} \end{aligned} \tag{2}$$

The values of the indicator variable are

$$\begin{aligned} &\mathbb{I}_1 = 0, && \text{without documentation} \\ &\mathbb{I}_2 = 1, && \text{with documentation.} \end{aligned} \tag{3}$$

The random effect γ_ℓ has distribution

$$\gamma_\ell | \sigma \sim \text{Normal}(0, \sigma), \qquad \ell = 1, \ldots, g, \qquad g \in \mathbb{N}. \tag{4}$$

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

$$\mathbb{P}\left(\text{Entry } l \text{ non-compliant}\right) = q_l = 1 - \prod_{ijk \text{ for lines in } l} [1 - p_{ijkl}], \tag{5}$$

where p_{ijkl} is the probability that the line with indices $ijkl$ is non-compliant, calculated from Eq. (1). Hence, for entries in container mode, we treat the entry as a Bernoulli random variable with probability defined by Eq. (5), while for entries in line mode, we treat each line as a Bernoulli random variable with probability as defined in Eq. (1).

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

3. Asymptotic analysis. We use asymptotic analysis to investigate how the precision of estimates depends on entry size, the number of entries, the probability of non-compliance and whether item types are mixed. This analysis comprises two parts. The first assumes that all items are a single type, which allows us to quantify how the amount of data, probability of non-compliance and entry size affect precision. The second part assumes that there are two different item types, and it explores how changing the proportion of entries with mixing both item types affects precision.

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

We estimate precision using an asymptotic estimate of the standard error. We calculate the precision from the square roots of the diagonal elements in the Fisher information matrix, $\mathcal{I}$, which is the expected value of the negative of the Hessian matrix of the log-likelihood evaluated at the value of the parameter.

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

$$\mathbb{P}(\text{entry compliant}) = (1-p)^S, \tag{6}$$

and the binomial log-likelihood for a set entry size S is $\log \mathcal{L}_S = I \log(1-(1-p)^S) + (N - I)\log((1-p)^S)$.

We estimate the standard error by calculating the Fisher information

$$\mathcal{I} = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}_S}{\partial p^2}\right] = -\frac{NS^2(1-p)^{S-2}}{(1-p)^S - 1}, \tag{7}$$

making the standard error

$$SE = \left(-\frac{NS^2(1-p)^{S-2}}{(1-p)^S - 1} \right)^{-1/2}. \tag{8}$$

Using Eq. (8) we can understand how the probability of non-compliance, the entry size and the number of lines affect the standard error, and we plot these relationships in Figure 1. The left plot shows that the standard error depends on the probability of non-compliance and that the relationship depends on the entry size. For all entry sizes, the standard error is small when the probability of non-compliance is small (below ~ 0.3). However, for larger values of the probability of non-compliance, the standard error increase significantly if the entry size is three or greater. The large p behaviour is driven by the $(1-p)^{S-2}$ term in Eq. (8), which means SE goes to 0 if $S = 1$, while it diverges if $S \geq 3$. Figure 1 also shows how the standard error decreases as the number of lines of data increases. The lower the entry size is, the lower the standard error, and, as $SE \sim \sqrt{1/N}$, container mode data with larger entry sizes require a large amount of data to reach the same standard error.

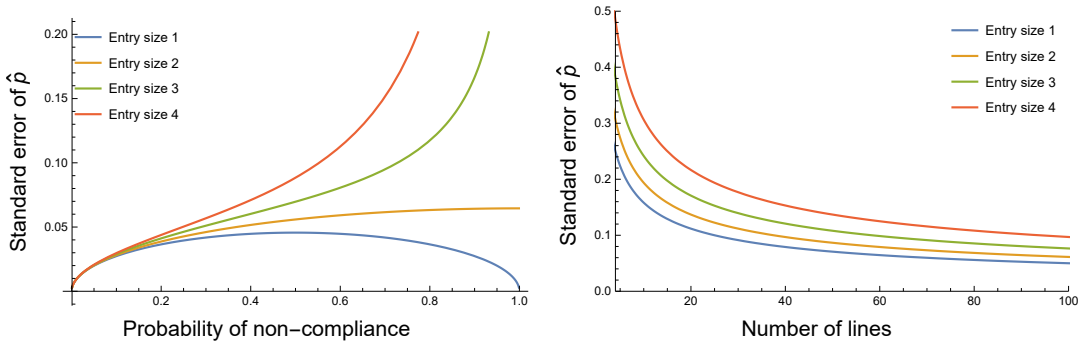

Fig 1: The standard error (Eq. (8)) as the probability of non-compliance, p, is varied (left) and as the number of lines are varied (right). For the left plot the number of lines is held constant at 120. For the right plot, the probability of non-compliance is held constant at 0.5

3.2. *Two item types.* In this section we consider a situation where there are two items with probabilities of non-compliance of p_1 and p_2, and we examine how these different items probabilities interact. We focus on a scenario where every entry is of size two, meaning there are three types of entries: only type 1; only type 2; or mixed, with one line of type 1 and 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 $S_1 = 0, S_2 = 2$. Then, the log-likelihood is

$$\log \mathcal{L} = \sum_{S_1,S_2} I(S_1, S_2) \log\left(1 - (1-p_1)^{S_1}(1-p_2)^{S_2}\right) + \\ (N(S_1, S_2) - I(S_1, S_2)) \log\left((1-p_1)^{S_1}(1-p_2)^{S_2}\right). \tag{9}$$

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

(10)
$$\frac{1}{2}\sqrt{-\frac{p_1\left(p_1^2-3p_1+2\right)\left(N_{1,2}(p_1-1)(p_2-2)p_2+4N_{2,2}(p_2-1)(p_1(p_2-1)-p_2)\right)}{N_{1,1}(p_1-1)(N_{1,2}(p_1-1)(p_2-2)p_2+4N_{2,2}(p_2-1)(p_1(p_2-1)-p_2))+N_{1,2}N_{2,2}(p_1-2)p_1(p_2-1)^2}},$$

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. (10) 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 0.75 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, while the proportion of mixed entries is below $\sim 90\%$.

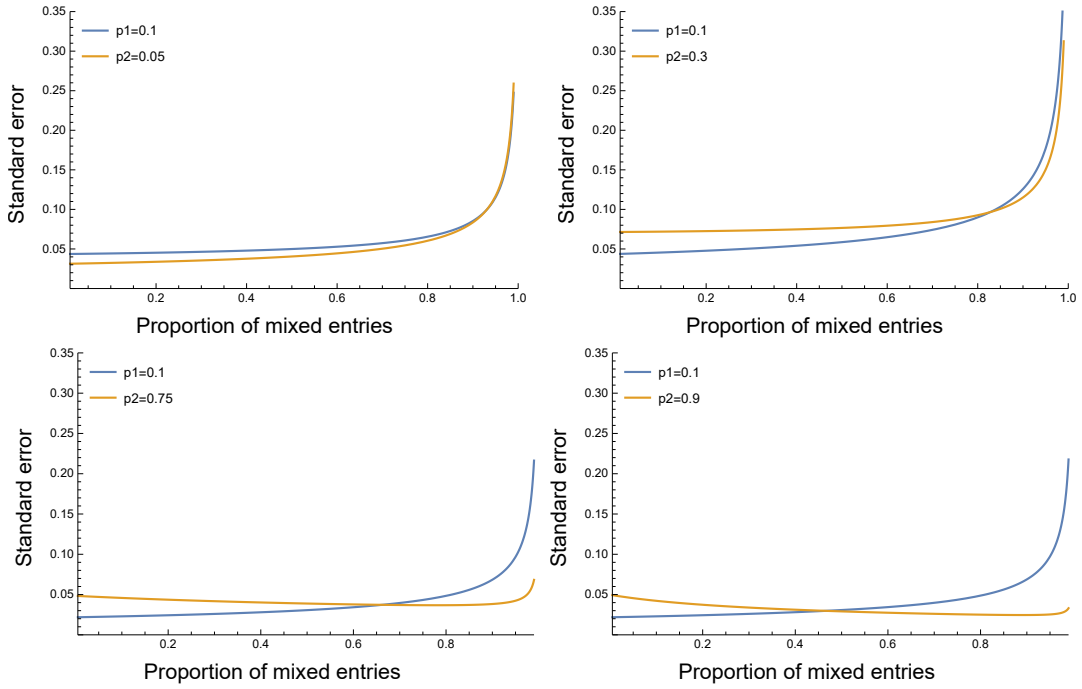

Fig 2: The standard error when for entry size 2, when the proportion of mixed entries is varied. In each plot there are 50 total entries and p_1 is held constant at 0.1, while p_2 goes from 0.05 up to 0.9.

3.3. *Asymptotic analysis conclusions.* From the asymptotic analysis we make two observations:

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

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

In this section, we generate 20 independent sets of data to fit our model to. We fit our model to different amounts of data, from 1,000 lines to 50,000 lines, for each for the 20 data sets separately, and we report the mean estimate, alongside the 90% CI across the 20 datasets. Thus, we can evaluate how we expect the precision of estimates to improve as the quantity of data increases.

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

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
⋮	⋮	⋮	⋮	⋮	⋮

4.2. *STAN model.* We model the system in a Bayesian framework, using the RSTAN package in R. We choose a Bayesian package due to the ease of specifying the model and that it could be updated, in principle, for any specification of the model for the probability of non-compliance. The basic structure for fitting the simulated data is to define the probability of non-compliance for each line (following Eq. (1))

$$\text{logit}(p_{ijk\ell}) = \alpha_i + \beta \mathbb{I}_j + \delta_k + \gamma_\ell, \tag{11}$$

where

$$\gamma_\ell | \sigma \sim \text{Normal}(0, \sigma), \quad \ell = 1, \ldots, g, \quad g \in \mathbb{N}. \tag{12}$$

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

$$\begin{aligned} \alpha_i &\sim \text{Normal}(-4, 4), \qquad i = 1, \ldots, a, \qquad a \in \mathbb{N}, \\ \beta &\sim \text{Normal}(0, 0.5), \\ \delta_k &\sim \text{Normal}(0, 0.5), \qquad k = 1, \ldots, d, \qquad d \in \mathbb{N} \\ \sigma &\sim \text{Normal}(0, 0.5) \end{aligned} \tag{13}$$

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

4.3. *Simulation results.* With sufficient data, we see that the STAN model estimates become close to the true value (Figure 3). However, for the α_i's, the spread of the estimates is higher for the smaller α_i's, and this spread reduces more slowly as the amount of data increases. It is not surprising that the estimates of the smaller values show poor performance, given our asymptotic analysis results (Section 3).

5. Case study. We apply our model to some real biosecurity data, to see how our understanding of the system changes by fitting the full data, compared to only using the container mode data. The data set is of furniture imports in 2020, and we break the data set into 52 weekly data sets for this analysis. Because container mode data with no non-compliance 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 four container mode entries with non-compliance, with the remaining weeks having fewer. Hence, we choose to analyse these three weeks as a demonstration of potential real-world differences between analysing all data and only line mode data. For confidentiality, for each week we re-name the item type, the country and the week to be integers (week 1, 2 and 3 do not correspond to the first 3 weeks of the year, and country 1 and item 1 are not the same in weeks 1 and 2). Relabeling the data does not cause issues because we are not trying to compare estimates between weeks or draw inferences between countries or item types in this analysis. A summary of the data is given in Table 2.

We fit our STAN model to the data from each week, using all the data and the line-only subset of the data. Because the line-only analysis is a subset of the data, not every country or item is present in the line-only data set. For the parameters that do match, we generate a scatter plot of the mean estimates to see how frequently we get different estimates (Figure 4). For many of the parameters, we get very similar results irrespective of which data we use.

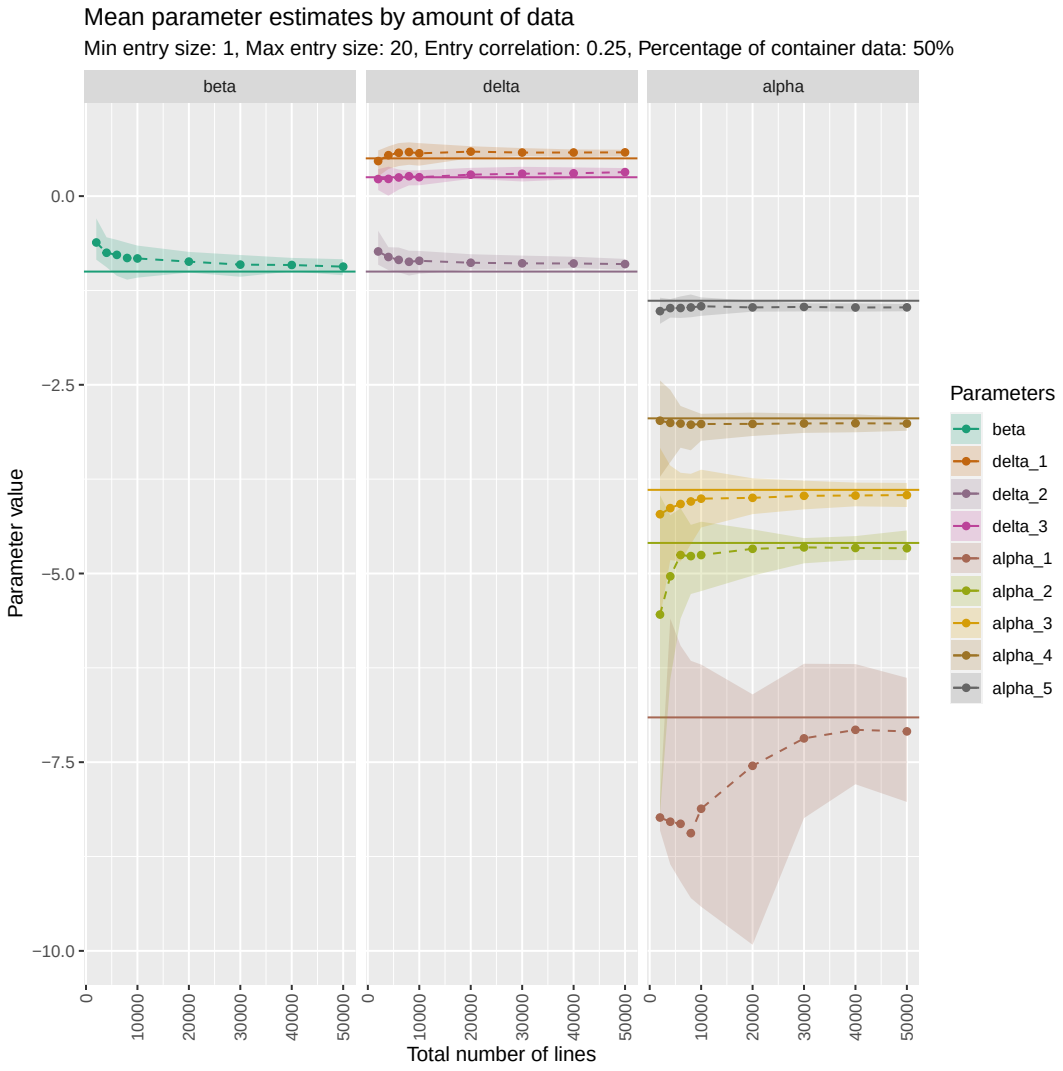

Fig 3: Results from the STAN model, fitted to simulated data. 50% of the data are in container mode and 50% are in line mode, and each entry is a random size between 1 and 20. Dots and dashed lines in each plot show the STAN estimates, while the solid lines show the true value of each parameter. The spread indicates the variation (90% confidence interval) between parameter estimates, over 20 repetitions of the simulated experiment.

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

This is not surprising, given that most of the data are in line mode and that line mode data gives more information than container mode data. However, it is interesting that including the container mode data results in quite large changes to some of the parameter estimates.

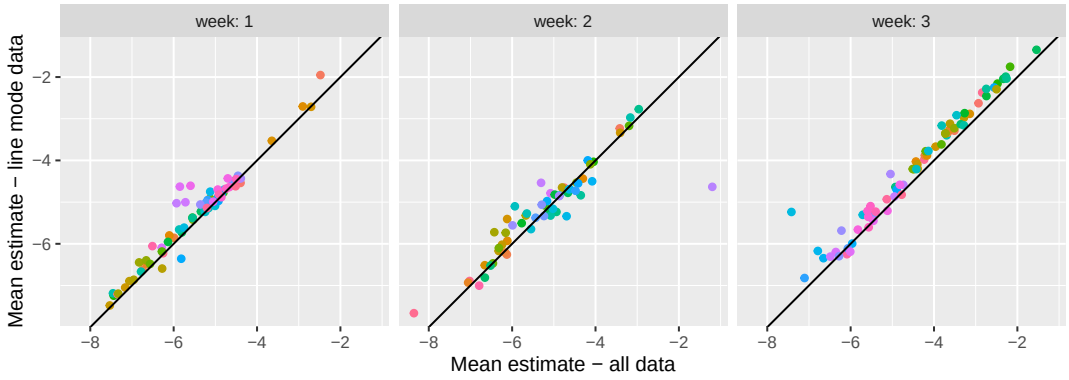

Fig 4: Scatter plot of parameter estimates using the line only data and all data. Dots that fall on the solid black line indicate when we get the same parameter estimate from each data set.

Conceivably, the container mode data might be drawn from a different distribution — that is, the varying parameter estimates represented in Figure 4 might represent a genuine signal. To investigate this possibility, we run the same analysis with a random subset of line mode entries set to container mode. As shown in Figure 5, the parameter estimates show to systematic error. We attribute the wider spread of values in Figure 5, as compared to Figure 4, to the smaller size of the underlying dataset.

Analysing the full data set, rather than the line-only data, gives us information about more items and countries. Table 3 shows all of the parameter estimates that we only get when including the container mode data in the analysis. The results of our case study demonstrate that, while parameter estimates based on container-mode data are subject to more uncertainty, their inclusion expands the scope of what can be studied.

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

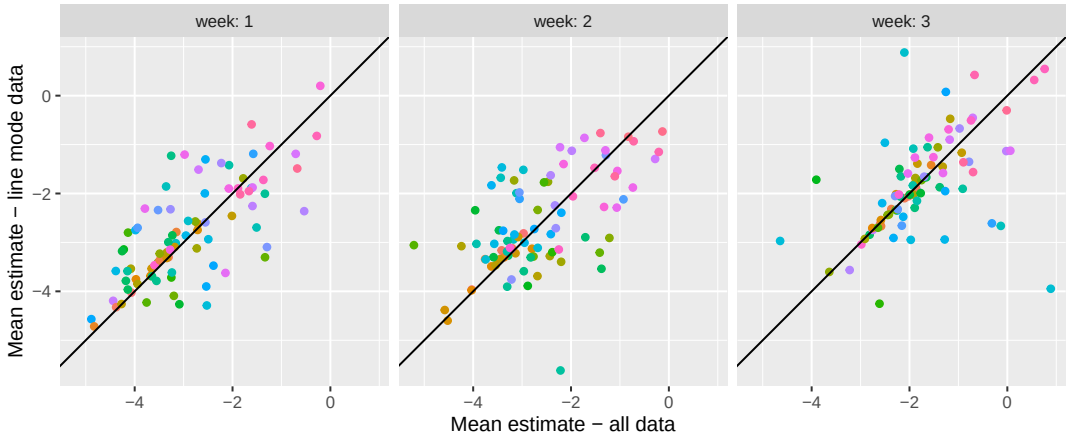

Fig 5: Scatter plot of parameter estimates using the line data only, with some entries set to container mode. The model was fit separately to the subset remaining in line mode and to the full set. The subset of entries in container mode was chosen randomly to match the proportion from the full data set.

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-compliant has divergent effects 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 of data where the presence of container mode data degrades model precision, compared to a line-only analysis.

While we are able to model the full dataset, the continuing presence of container mode data will be a barrier for future data analysis. Most statistical and machine learning algorithms are designed such that each row of data will have explanatory variables and an outcome. With container mode data, we have the explanatory variables, but only partial information about the

outcome. Hence, we need customised algorithms (such as described in this paper) to analyse it appropriately. Even so, many simplifying assumptions have been made: for example, that a given item's probability of non-compliance does not vary over time. These simplifications do not change our conclusions, as the analysis still serves to highlight the differences between container and line mode. However, without proper care, the specific results of our case-study analysis should not be applied to decision making.

The situation in our biosecurity case study is representative of a wider issue in applied statistics, where a portion of the available data is not in a form able to be analysed in the usual way. In the context of missing data it is widely accepted that methods for data imputation are worthwhile and provide better results than ignoring or deleting missing data (Nakagawa and Freckleton, 2008; Ren et al., 2023). However, for the partially-pooled case that we study, there is not a range of off-the-shelf methodologies to incorporate all of the data into a single analysis, and the simplest way to analyse these data using machine learning software is to ignore the container-mode lines. The differences between the results of the case study when we use only the line-mode data compared to when we use all data highlights the utility of developing models that are capable of using all of the data.

While in our case the pooling of outcomes in container mode is an artifact of the system, pooled testing are often designed strategically in public health to improve efficiency. The strain on the PCR testing system has prompted regimes to identify cases while minimising the rounds of testing Mutesa et al. (2021). While it is clear that the way pooling is conducted affects how quickly infections can be identified, in some circumstances, the outcomes of pooled testing can also be used to estimate parameters, and our work has implications for these situations Delaigle and Hall (2015); Chatterjee and Bandyopadhyay (2020); McMahan et al. (2017); Liu et al. (2020). From our asymptotic analysis, it is clear that even when only estimating a prevalence, for a fixed number of tests, the standard error of the estimate depends on pool size and therefore there would be an optimal pool size, which would depend on the prevalence Brynildsrud (2020). When there is a model with parameters being estimated, the added complexity could compound the issue, making pool sizing a more important aspect of study design. Furthermore, as we demonstrate here, it is not only pool size that matters, it is how samples with different characteristics are grouped together. Hence, for a given study, there may not be one optimal pooling strategy, and there would likely be trade-offs between different parameters when it comes to maximising the precision of estimates.

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

REFERENCES

BAKER, C. M. and BODE, M. (2020). Recent advances of quantitative modeling to support invasive species eradication on islands. *Conservation Science and Practice* **n/a** e246. _eprint: https://conbio.onlinelibrary.wiley.com/doi/pdf/10.1111/csp2.246. https://doi.org/10.1111/csp2.246

BINNY, R. N., INNES, J., FITZGERALD, N., PECH, R., JAMES, A., PRICE, R., GILLIES, C. and BYROM, A. E. (2021). Long-term biodiversity trajectories for pest-managed ecological restorations: eradication vs. suppression. *Ecological Monographs* **n/a**. _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/ecm.1439. https://doi.org/10.1002/ecm.1439

BROOK, L. A., JOHNSON, C. N. and RITCHIE, E. G. (2012). Effects of predator control on behaviour of an apex predator and indirect consequences for mesopredator suppression. *Journal of Applied Ecology* **49** 1278–1286. https://doi.org/10.1111/j.1365-2664.2012.02207.x

BROOKS, S., GELMAN, A., JONES, G. and MEGN, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, New York. https://doi.org/10.1201/b10905

BRYNILDSRUD, O. (2020). COVID-19 prevalence estimation by random sampling in population - optimal sample pooling under varying assumptions about true prevalence. *BMC Medical Research Methodology* **20** 196. https://doi.org/10.1186/s12874-020-01081-0

CHATTERJEE, A. and BANDYOPADHYAY, T. (2020). Regression models for group testing: Identifiability and asymptotics. *Journal of Statistical Planning and Inference* **204** 141–152. https://doi.org/10.1016/j.jspi.2019.05.003

DELAIGLE, A. and HALL, P. (2015). Nonparametric methods for group testing data, taking dilution into account. *Biometrika* **102** 871–887. https://doi.org/10.1093/biomet/asv049

HELMSTEDT, K. J., SHAW, J. D., BODE, M., TERAUDS, A., SPRINGER, K., ROBINSON, S. A. and POSSINGHAM, H. P. (2016). Prioritizing eradication actions on islands: it's not all or nothing. *Journal of Applied Ecology* **53** 733–741. https://doi.org/10.1111/1365-2664.12599

HEPWORTH, G. and BIGGERSTAFF, B. J. (2017). Bias Correction in Estimating Proportions by Pooled Testing. *Journal of Agricultural, Biological and Environmental Statistics* **22** 602–614. https://doi.org/10.1007/s13253-017-0297-2

HEPWORTH, G. and BIGGERSTAFF, B. J. (2021). Bias Correction in Estimating Proportions by Imperfect Pooled Testing. *Journal of Agricultural, Biological and Environmental Statistics* **26** 90–104. https://doi.org/10.1007/s13253-020-00411-5

HEPWORTH, G. and WATSON, R. (2009). Debiased estimation of proportions in group testing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **58** 105–121. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9876.2008.00639.x. https://doi.org/10.1111/j.1467-9876.2008.00639.x

HOLMES, N. D., SPATZ, D. R., OPPEL, S., TERSHY, B., CROLL, D. A., KEITT, B., GENOVESI, P., BURFIELD, I. J., WILL, D. J., BOND, A. L., WEGMANN, A., AGUIRRE-MUÑOZ, A., RAINE, A. F., KNAPP, C. R., HUNG, C.-H., WINGATE, D., HAGEN, E., MÉNDEZ-SÁNCHEZ, F., ROCAMORA, G., YUAN, H.-W., FRIC, J., MILLETT, J., RUSSELL, J., LISKE-CLARK, J., VIDAL, E., JOURDAN, H., CAMPBELL, K., SPRINGER, K., SWINNERTON, K., GIBBONS-DECHERONG, L., LANGRAND, O., BROOKE, M. D. L., MCMINN, M., BUNBURY, N., OLIVEIRA, N., SPOSIMO, P., GERALDES, P., MCCLELLAND, P., HODUM, P., RYAN, P. G., BORROTO-PÁEZ, R., PIERCE, R., GRIFFITHS, R., FISHER, R. N., WANLESS, R., PASACHNIK, S. A., CRANWELL, S., MICOL, T. and BUTCHART, S. H. M. (2019). Globally important islands where eradicating invasive mammals will benefit highly threatened vertebrates. *PLOS ONE* **14** e0212128. https://doi.org/10.1371/journal.pone.0212128

JARDINE, S. L. and SANCHIRICO, J. N. (2018). Estimating the cost of invasive species control. *Journal of Environmental Economics and Management* **87** 242–257. https://doi.org/10.1016/j.jeem.2017.07.004

KUMAR RAI, P. and SINGH, J. S. (2020). Invasive alien plant species: Their impact on environment, ecosystem services and human health. *Ecological Indicators* **111** 106020. https://doi.org/10.1016/j.ecolind.2019.106020

LIU, Y., MCMAHAN, C. S., TEBBS, J. M., GALLAGHER, C. M. and BILDER, C. R. (2020). Generalized additive regression for group testing data. *Biostatistics* **kxaa003**. https://doi.org/10.1093/biostatistics/kxaa003

MCMAHAN, C. S., TEBBS, J. M., HANSON, T. E. and BILDER, C. R. (2017). Bayesian regression for group testing data. *Biometrics* **73** 1443–1452. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12704. https://doi.org/10.1111/biom.12704

MUTESA, L., NDISHIMYE, P., BUTERA, Y., SOUOPGUI, J., UWINEZA, A., RUTAYISIRE, R., NDORICIMPAYE, E. L., MUSONI, E., RUJENI, N., NYATANYI, T., NTAGWABIRA, E., SEMAKULA, M., MUSANABAGANWA, C., NYAMWASA, D., NDASHIMYE, M., UJENEZA, E., MWIKARAGO, I. E., MUVUNYI, C. M., MAZARATI, J. B., NSANZIMANA, S., TUROK, N. and NDIFON, W. (2021). A pooled testing strategy for

identifying SARS-CoV-2 at low prevalence. *Nature* **589** 276–280. Number: 7841 Publisher: Nature Publishing Group. https://doi.org/10.1038/s41586-020-2885-5

NAKAGAWA, S. and FRECKLETON, R. P. (2008). Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution* **23** 592–596. Publisher: Elsevier. https://doi.org/10.1016/j.tree.2008.06.014

REN, L., WANG, T., SEKHARI SEKLOULI, A., ZHANG, H. and BOURAS, A. (2023). A review on missing values for main challenges and methods. *Information Systems* 102268. https://doi.org/10.1016/j.is.2023.102268

SCHAARSCHMIDT, F. (2007). Experimental design for one-sided confidence intervals or hypothesis tests in binomial group testing. *Communications in Biometry and Crop Science*. Publisher: Faculty of Agriculture and Biology, Warsaw Agricultural University, Poland.

SHAROV, A. A., LEONARD, D., LIEBHOLD, A. M., ROBERTS, E. A. and DICKERSON, W. (2002). "Slow The Spread": A National Program to Contain the Gypsy Moth. *Journal of Forestry* **100** 30–36.

THOMPSON, K. H. (1962). Estimation of the Proportion of Vectors in a Natural Population of Insects. *Biometrics* **18** 568–578. Publisher: [Wiley, International Biometric Society]. https://doi.org/10.2307/2527902

WENGER, A. S., ADAMS, V. M., IACONA, G. D., LOHR, C., PRESSEY, R. L., MORRIS, K. and CRAIGIE, I. D. (2017). Estimating realistic costs for strategic management planning of invasive species eradications on islands. *Biological Invasions* 1–19. https://doi.org/10.1007/s10530-017-1627-6