Statistical profiling to predict the biosecurity risk presented by non-compliant international passengers

Stephen E Lane[1], Richard Gao[2], Matthew Chisholm[2], and Andrew P Robinson[1]

[1] *Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne, Parkville, Victoria 3010, Australia,* `lane.s@unimelb.edu.au`

[2] *Department of Agriculture and Water Resources, Canberra, Australian Capital Territory 2601, Australia*

February 15, 2017

Abstract

Biosecurity risk material (BRM) presents a clear and significant threat to national and international environmental and economic assets. Intercepting BRM carried by non-compliant international passengers is a key priority of border biosecurity services. Global travel rates are constantly increasing, which complicates this important responsibility, and necessitates judicious intervention. Selection of passengers for intervention is generally performed manually, and the quality of the selection depends on the experience and judgement of the officer making the selection. In this article we report on a case study to assess the predictive ability of statistical profiling methods that predict non-compliance with biosecurity regulations using data obtained from regulatory documents as inputs. We then evaluate the performance arising from using risk predictions to select higher risk passengers for screening. We find that both prediction performance and screening higher risk passengers from regulatory documents are superior to manual and random screening, and recommend that authorities further investigate statistical profiling for efficient intervention of biosecurity risk material on incoming passengers.

Keywords: Biosecurity, risk prediction, profiling, international air passengers

1 Introduction

Increasing globalisation and international travel increase the risks of pests and diseases invading novel areas and systems (e.g., Hulme, 2009) . Passengers arriving at international borders represent biosecurity risk because they may carry pests and/or diseases that are transmitted by biosecurity risk material (BRM), for example fruit from the originating country may be infested by fruit flies, and uncooked meat articles may be carrying Foot and Mouth Disease (FMD).

The impact of invasive pests and diseases is economically and environmentally substantial. The annual cost of invasive weeds in Australia has been estimated as AUD$4bn (Sinden et al., 2005), and of invasive species generally to over USD$200bn (Pimentel, 2011) annually. Furthermore, invasions can lead to a loss of ecosystem diversity — for example, invasive alien species were estimated to decrease local diversity by 51% (Vilà et al., 2011). Invasives are cited by the IUCN as being the most significant threat to biodiversity in islands, and the second most significant after climate change elsewhere (IUCN, nodate).

The Department of Agriculture and Water Resources (the department) has primary responsibility for managing the biosecurity system in Australia (Department of Agriculture and Water Resources, nodate). The department undertakes interventions of various kinds in the many pathways over which it has regulatory authority, based on the principle that prevention is better than cure (e.g., Leung et al., 2002). For example, the department screens incoming passengers at clearance points (e.g., sea and airports) for BRM. This screening process typically involves assessment of the passenger's paperwork, an interview, and possibly examination of the passenger's effects by x-ray or a Detector Dog Unit, or physical inspection. The screening process is focused on detecting undeclared BRM only; declared BRM can be intercepted without screening, because it is voluntarily presented for inspection.

Screening all international passengers with equal effort is inefficient because different cohorts of passengers present different risks. Profiling of some form is generally involved in screening passengers for BRM both in Australia and internationally. Profiling is generally performed manually, whereby the officer assesses passengers to decide on how they should be processed. The quality of this assessment depends on the experience and judgement of the officer.

Increasing numbers of passengers and greater diversity in originating countries passing through Australia's borders mean that the profiling task for officers is becoming more complex. For example, the number of international passenger clearances increased by 4% from 2013–14 to 2014–15 and 6% from 2014–15 to 2015–16 (Department of Agriculture and Water Resources, 2016). If passengers can be classified into groups or cohorts that reflect varying degrees of risk of non-compliance, then focusing intervention efforts upon those with higher risk should result in a higher rate of detection of non-compliance, whilst at the same time expending the same or fewer resources (R. M. Cannon, 2009).

Expert judgement for allocating passengers into groups of varying non-compliance risk can be time consuming and may be inefficient (Burgman, 2016). An alternative approach is to augment the allocation process by using a statistical profiling tool. Given a cohort of arriving passengers and associated traits/demographics, along with a risk prediction rule that places these passengers into various risk groups, a selection strategy can be defined to maximise detection of non-compliance.

Research into risk-based sampling for border intervention has historically focused on goods and commodities entering through ports. For example, Robinson, Burgman, and R. Cannon (2011) investigated risk-based allocation of inspection resources for unit load devices (air transportation containers), and Hua, Li, and Tao (2005) used a combination of logistic regression and clustering to define a rule-based risk decision system for inspection of goods entering China. We are not aware of any reported research into the performance of statistical profiling for risk-based selection of arriving passengers for the purpose of detecting BRM. Melo et al. (2014) investigated associations of air passenger traits with possession of BRM, but did not look further at the possibilities of using this information for profiling. Related but more specific investigations were undertaken by Lin et al. (2009), who constructed a model to predict FMD status in meat illegally carried by air passengers, and Lai, Hwang, and Chou (2012) who constructed a similar model for avian influenza virus. Shih, Chou, and Morley (2005) investigated associations of passenger traits with monthly non-compliance counts, but did not investigate risk prediction for individual passengers.

We have two aims for the present article: (i) to assess the predictive ability of statistical profiling methods for predicting non-compliance with biosecurity regulations using passenger traits recorded on regulated

documentation such as Incoming Passenger Cards (IPCs); and (ii) to evaluate the performance of the risk predictions in terms of screening higher risk passengers. We performed assessments via a cross-validation study using inspection data, which we describe in Sections 2 (data) and 3 (methods). We present the results of the study in Section 4, and make closing remarks in Section 5.

2 Data

Staff at Kingsford-Smith International Airport in Sydney, Australia conducted a census operation on a particular arriving flight between 18 June and 20 August 2015 inclusive. The flight was chosen due to a relatively high incidence of BRM interceptions, and a desire to minimise unnecessary screening activity by improving profile specificity. We used the data collected during this operation to perform a detailed analysis of passenger non-compliance. During normal operations, many passengers are not subjected to inspection, and passenger details are only recorded when undeclared biosecurity risk material (BRM) is found.

During the census operation all passengers who arrived were screened. Screening in this context refers to passenger's luggage being examined, e.g., by X-ray and/or opening and physically examining the luggage contents. Passenger traits were recorded for all passengers, along with details of any BRM located; passengers found with undeclared BRM were labelled as being non-compliant. The dataset comprised 3361 records, of which 6.5% were non-compliant. Table 1 shows the passenger traits that were recorded. Cross-classification tables cannot be published for privacy reasons.

Table 1: Description of passenger traits recorded during the screening process. Stage 1 predictor variables are those that can be used for profiling without direct reference to the IPC, whilst Stage 2 predictor variables require the IPC. See Section 3.1 for modelling details.

Non-compliance status	A binary variable, with value 1 for a non-compliant result and 0 for a compliant result in the screening process (where 'non-compliant' indicates the presence of any undeclared BRM).
Stage 1 predictor variables	
Year of birth/age	A numerical variable recording the passenger's year of birth. For ease of interpretation, this was transformed to age at the time of screening (years).
Sex	A binary variable, with values *male* and *female*.
Stage 2 predictor variables	
Citizenship group	A categorical variable indicating the passenger's citizenship, as found on the IPC, and grouped into geographical regions.
Declaration status	A binary variable, with value 1 for a declarant and 0 for a non-declarant, based on the passenger's Incoming Passenger Card (IPC).
Occupation	A categorical variable for the passenger's occupation, as recorded on the IPC.
Visit reason	A categorical variable for the passenger's visit reason, as found on the IPC.

2.1 Data preprocessing

Some processing of the dataset was required before model fitting. Some levels within passenger traits were sufficiently rarely observed that we decided to collapse these into a *not otherwise specified* level; specifically, levels that contained less than 50 observations were collapsed together prior to formal statistical modelling.

3 Methods

Repeated tenfold cross-validation was used for model comparison in this study, with ten repeats, resulting in 100 training/testing datasets for comparison.

3.1 Prediction models compared

For each training/testing dataset, the following models were fit. Brief descriptions are provided below, with more detail in Appendix A.1.

1. semi-parametric logistic regression (**GAM**)
2. a random forest with the number of trees optimised (**RF-caret**)
3. a gradient boosting machine (GBM) with custom feature selection/collapsing and tuning parameter selection (**GBM-custom**)
4. a GBM with optimised tuning parameters (interaction depth, number of trees/base learners, learning rate; **GBM-caret**)
5. a neural network with the number of units in the hidden layer, and the weight decay optimised (**NN-caret**)
6. a Bayesian (shrinkage) logistic regression models with normal priors (**Bayes-normal**)
7. a Bayesian lasso model with Laplace priors (**Bayes-lasso**)

The random forest (Breiman, 2001), GBM (Friedman, 2001) and neural network (B. D. Ripley, 2008) models were chosen due to their popularity in machine learning, ease of implementation in R, and well-documented performance (e.g., Hastie, Tibshirani, and Jerome Friedman, 2009).

Random forests are a machine-learning technique that perform *ensemble learning* for classification, where the ensemble consists of a large number of classification trees (known as base learners). Each base learner is trained on a random sample of the data, and is grown to the maximum possible extent. Growing the base learner to the maximum possible extent results in overfitting, where predictions on a new dataset perform poorly, hence the resampling and refitting of multiple base learners. To make a prediction for a new observation, each base learner classifies that observation. The classification that occurs the most (over all trees in the forest) is taken to be the prediction for the observation.

The GBM also provides ensemble learning, but the base learners in a GBM are *weak* learners; they are not grown to the maximum possible extent. The GBM starts with an imperfect model for the data (i.e. the base learner that is not grown maximally), and constructs a new model by successively fitting the residuals of the current model, using the same class of base learners as the initial imperfect model.

Neural network (NN) models are inspired by biological networks within the brain, in which messages are passed between neurons to send instructions for various biological processes. In NN machine learning, the *messages* are the weighted predictor variables, which are passed along the network until they reach the output variables. Due to the many paths a message can take through the network, NN models provide a very flexible way of non-linear model fitting.

Machine learning algorithms require custom programming to be implemented in most operational environments, so we decided to also compare logistic regression (with a semiparametric model, e.g. Simon N Wood, 2011), which can be more readily implemented.

The Bayesian shrinkage methods (Park and Casella, 2008) were chosen as a compromise between the machine learning and logistic regression methods. With some passenger traits being rarely observed (Section 2), we felt these models would provide robustness against any separation issues that may occur.

Further, for each method we fit models with two different passenger trait sets. In the first stage, we use passenger traits that could be identified manually by the officer, without reference to the IPC. Profiles formed from this grouping could be used quickly by officers in the facility to select passengers for further screening. Models in this stage use age and as predictors, shown under the sub-heading *Stage 1 predictor variables* in Table 1. The second stage uses the first stage passenger traits, along with those listed under the sub-heading *Stage 2 predictor variables* as given in Table 1. The extra passenger traits, citizenship

group, declaration status, occupation and visit reason all require inspection of IPCs, and are thus more time consuming. The second stage models possibly provide better predictions however, so we compare both stages for each model.

3.2 Prediction metrics

3.2.1 Overall model comparison

Overall model predictive ability was compared using the area under the receiver operator curve (AUC; e.g. Hanley and McNeil, 1982), and predictive log-loss (e.g. Winkler, 1967). AUC is the curve created by plotting the true positive rate against the false positive rate (in the testing dataset) at various threshold probabilities, as if every passenger in the testing dataset were screened for non-compliance. Larger AUC values are associated with models that have better predictive ability.

Predictive log-loss measures how well the probabilities of non-compliance are predicted[1]. Given a sample of n passengers, let $y_i = 1$ if a passenger is non-compliant and 0 otherwise, and $\hat{p}_i$ be the predicted probability that a passenger is non-compliant. Then, predictive log-loss is defined as $-(1/n)\sum[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)]$. The log penalises 'confident' predictions (close to 0 or 1) that are incorrect. Smaller values of predictive log-loss are associated with models that produce better predictions.

3.2.2 Passenger screening comparisons

In practice, it is prohibitively expensive and slow to inspect every passenger, and so some form of screening selection is required. As described earlier, manual profiling is performed by officers who select passengers for screening. We simulated a targeted screening strategy that uses the passenger's predicted probabilities of non-compliance to select those that are to be screened. We perform this screening step on the withheld testing datasets described above. In this targeted strategy, passengers are ordered according to predicted risk of non-compliance with the top $P_{\text{scr}}\%$ of passengers all screened. In practice, we would also randomly select a proportion of the remaining passengers to keep current information about apparently low-risk cohorts; we omit this sample in the current exercise.

Some commonly reported metrics for these classification tasks are the positive predictive value, which is the proportion of selected passengers who were truly non-compliant, and the true positive rate, which is the proportion of non-compliant passengers selected. It is however, informative to consider the results in terms of *efficiency*, which is the value of, for example, positive predictive value at a fixed screening proportion, relative to what would be achieved under a random screening approach. This approach has the advantage that the various metrics give equivalent estimated efficiency, simplifying the comparison between methods. Further, we can compare these efficiency estimates to those estimated for manual profiling from previous data, as we now describe.

As part of performance measurement, the Department regularly conducts post-screening inspection surveys to estimate the efficiency of the manual profiling by officers[2]. Results from the most recent survey data show that the estimated non-compliance rate during the current census was less than that estimated by the survey. However, if we assume that the relative efficiency of manual profiling to the overall non-compliance rate (equivalent to random profiling, as discussed above) is the same in the current data collection as it was during the survey, then we can compare manual profiling performance to statistical profiling performance via efficiency. During the recently conducted survey, manual profiling by officers resulted in an estimated efficiency of 1.3 as compared with random sampling.

[1] In contrast to AUC, which measures the effect of classification at various thresholds.

[2] We are unable to show this data due to privacy reasons.

4 Results

4.1 Overall predictive ability

Figure 1 displays boxplots of the AUC and predictive log-loss for all methods. The performance of the **GBM-custom** and **RF-caret** were extremely poor in comparison to the other methods based on predictive log-loss, so we do not consider these methods any further. There is little to separate models fit using Stage 1 or Stage 1 and 2 passenger traits based on predictive log-loss. Comparisons with AUC however show that including all passenger traits in the models improves predictive ability. Similarly, there is little difference between methods, with **GBM-caret** having the highest (although not statistically significantly so) median AUC and lowest log-loss.

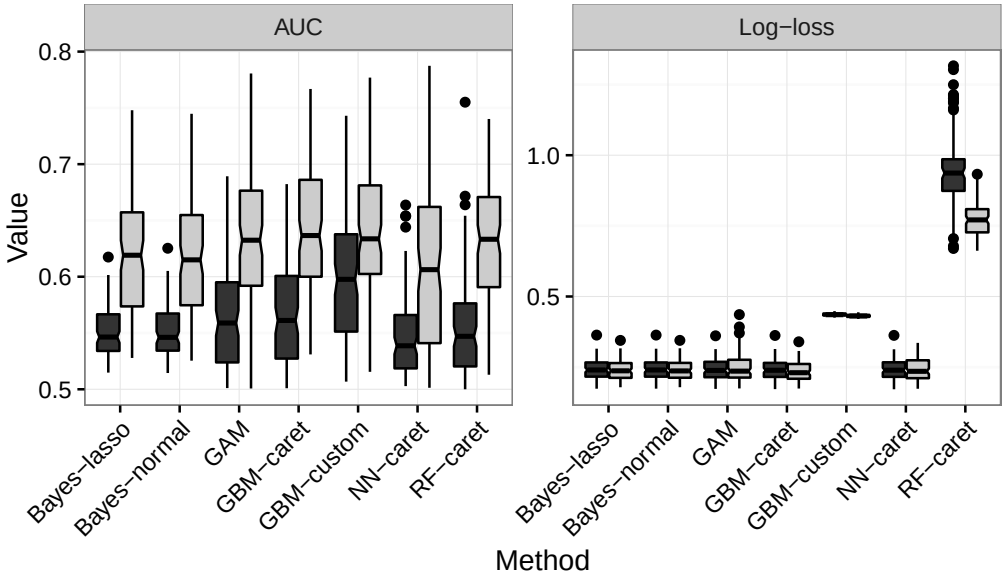

Figure 1: Comparison of AUC and predictive log-loss between methods. Dark grey fill are from models using Stage 1 passenger traits, whilst light grey fill are from models using Stage 1 and 2 passenger traits.

4.2 Passenger screening performance

Figure 2 shows efficiency as a function of screening rate. Panels in the figure represent the prediction method, and the solid black line is from models using Stage 1 passenger traits whilst the dashed black line is from models using Stage 1 and 2 passenger traits. Models fitted using Stage 1 and 2 passenger traits outperform those fitted using only Stage 1 passenger traits for all methods except **NN-caret** from very low screening rates, until approximately 60–70% of passengers are screened.

Also shown in Figure 2 is the efficiency of manual profiling by officers (1.3), as discussed in Section 3.2.2. It is clear from this figure that profiles created from statistical modelling (using Stage 1 and 2 passenger traits) outperform manual profiling by up to 50% in efficiency.

4.3 Importance of passenger traits

To assess the importance of individual passenger traits, we fit a model to the full dataset using both Stage 1 and 2 passenger traits. We chose to use the **GBM-caret** method due to its better performance as demonstrated[3] by Figures 1 and 2.

[3]Whilst not significantly superior, it has the benefits of flexibility and ease of implementation as discussed in Section 3.1.

Figure 3 displays the relative importance of the passenger traits. Importance as calculated in the `GBM` package is the relative influence a passenger trait has on reducing the loss function used in the GBM. Importance was standardised so the sum of all estimated passenger trait influences equalled 100. Age, occupation and the reason for the visit are estimated as being much more important for predicting non-compliance than a passenger's declaration status, citizenship group or sex.

Figure 4 displays the marginal effect of the three passenger traits[4] with the largest relative importance (Figure 3); marginal effects being calculated using the tree traversal method (Friedman, 2001). Age has a non-linear effect on non-compliance, with large relative jumps at around 20 and 45 years of age, and a large fall at around 60 years of age. Also shown in the occupation and visit reason panels are the number of passengers with each level of the trait, which provides a gauge of the precision related to the estimated marginal effect.

A demonstration of the ability of the GBM method to pick up interaction effects is shown in the bottom right panel of Figure 4. This panel shows the interaction effect of occupation and citizenship grouping; citizenship groupings are represented by different line types. The effect of occupation in passengers who are citizens of country Y is clearly different to that of passengers from the other two countries of citizenship.

[4] Levels of the traits have been masked due to privacy reasons.

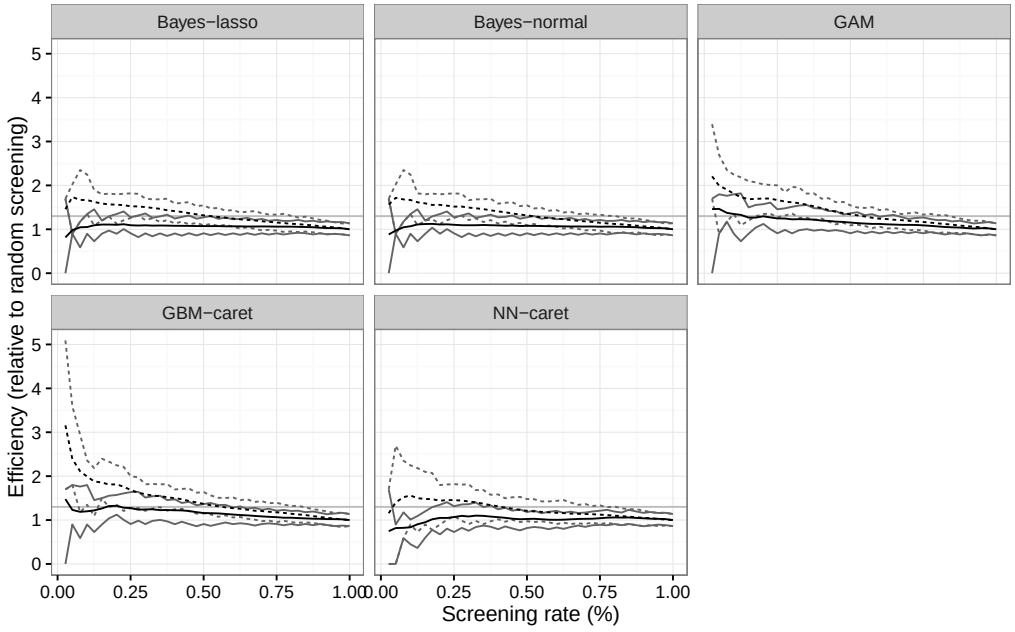

Figure 2: Comparison of efficiency as a function of screening rate for each method. The solid black line is from models using Stage 1 passenger traits, and the dashed black line is from models using Stage 1 and 2 passenger traits. 50% pointwise intervals are shown in grey. The horizontal grey line shows the estimated efficiency of manual profiling.

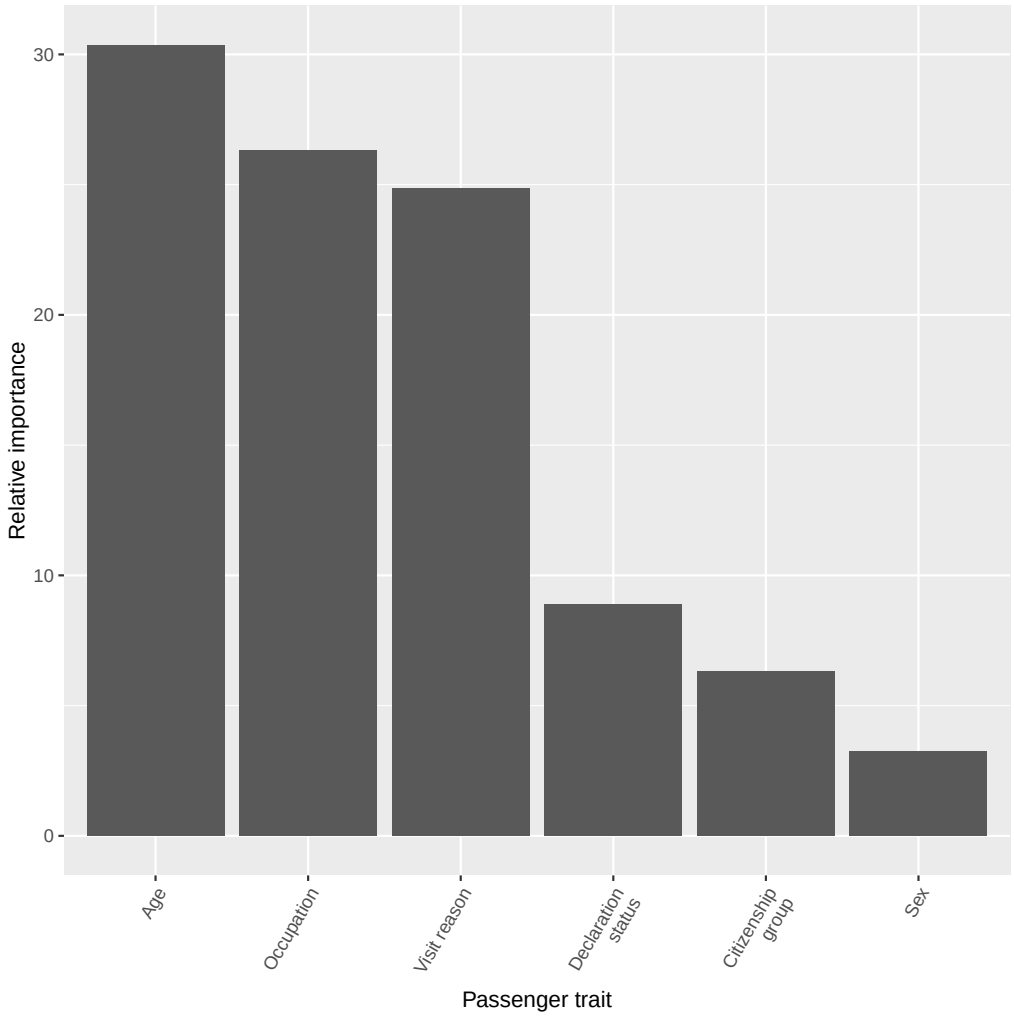

Figure 3: Relative importance of passenger traits in predicting biosecurity risk material non-compliance in a GBM model fit to the full dataset.

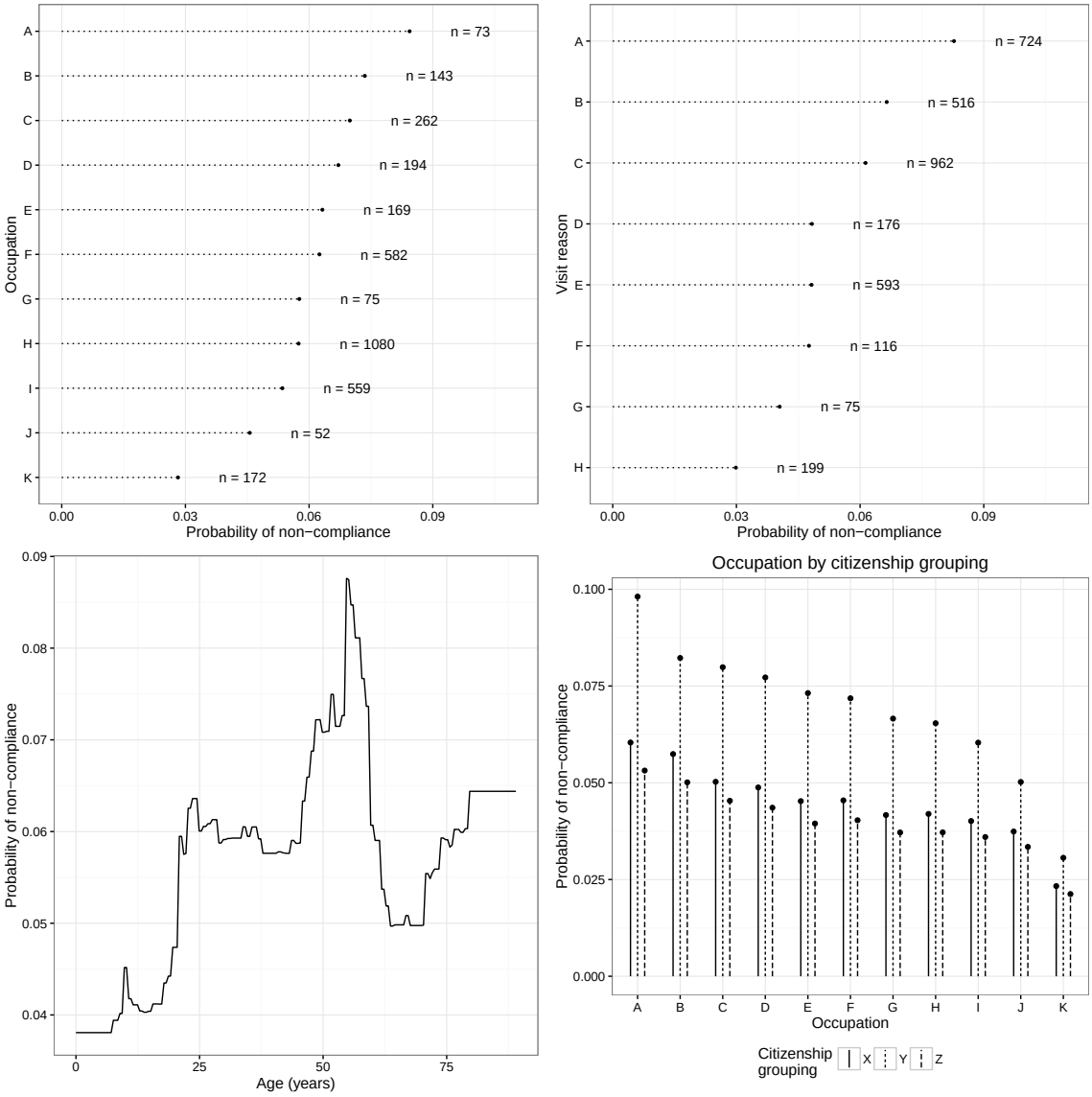

Figure 4: Marginal effect of selected passenger traits in predicting biosecurity risk material non-compliance in a GBM model fit to the full dataset.

5 Discussion

The results of the present case study show that for this dataset — using only information collected from IPCs — detection of non-compliance can be improved using statistical profiling. Importantly, we demonstrate this result holds relative to both random and manual selection. The predictive ability of most methods trialled was good, despite having a small dataset for modelling (both in terms of number of observations and variables). We conclude that data collected from IPCs is sufficient for constructing beneficial models that predict non-compliance for a wide range of model-fitting methods.

For any reasonable screening rate (up to 50% screening of all passengers), the statistical profiling methods were more efficient than manual profiling (Figure 2). This observation is important in the context of increasing passenger clearances as discussed in the introduction. With the increase in passenger clearances, either resources are spread more thinly on the ground, resulting in a lower screening rate, and/or budgets need to be increased for the extra personnel required to maintain the status quo. The results of the present case study indicate that it may in fact be possible to decrease the screening rate, whilst still maintaining a suitable level of biosecurity risk material interceptions. The end result of such a change, were it to be implemented, is likely to be a more efficient use of ground staff to meet the currently challenging environment.

In situations where automated screening — that is, electronic screening prior to entry — is not possible, it would be advantageous to have approximate profiling rules that could be used based upon a visual profile only. We fit the models in two stages against this contingency. The first stage used passenger traits amenable to manual decision making: age and sex, whilst the second stage added in variables that would only be available upon inspection of the IPCs. Models fit using all available data from the IPCs (Stage 2 models) clearly outperformed the Stage 1 models (Figure 1). We thus recommend that profiles be created using all available data, and be automated if possible.

As demonstrated in Figure 2 however, the Stage 1 models were still more efficient than current manual profiling, up to screening rates of approximately 30% using the **GBM-caret** method. These results suggest that if automated processing is not possible in the near future, profiles can still be constructed from Stage 1 models — for example by creating simple scoring rules based on effect sizes and variation, similar to the nomograms created in the medical literature for survival prediction (e.g., Callegaro et al., 2016). Such simple scoring rules still require time to process and if to be implemented via a visual screening approach, may be too cognitively demanding. However the benefit of doing so would be that direct officer engagement with passengers is lowered, as no interrogation of IPCs is required. Again, this has the flow-on effect of efficiency. Until such a time that information provided on IPCs is collected electronically, and likewise screening is automated, this approach should be trialled further.

This analysis enabled testing of a long-held informal belief among biosecurity inspectors that the declaration status of a passenger did not necessarily indicate their compliance status. That is, the declaration of one type of BRM does not reliably preclude the possibility that other BRM may be present. If declarant passengers presented lower risk than non-declarant passengers then we would expect declaration status to play an important role in the model. Figure 3 shows, however, that declaration status was not as important as age, occupation or visit reason in predicting non-compliance. We conclude that declarant passengers may still present risk of undetected BRM, with the caveat of course that the declaration of BRM still reduces the material risk.

The case study presented was restricted to only one dataset, collected during a fixed time period of two months, and was based upon a single flight number arriving at a single airport. Creating prediction rules to be used operationally was not the focus, and further research should investigate multiple flights covering different arrival times, and multiple entry points. Country/port of origin of the passenger, and time of departure/arrival could be investigated in expanded datasets. However, the clear performance advantages of statistical profiling compared with manual profiling as demonstrated in this study are only likely to improve with richer, expanded datasets.

Finally, in order to maintain current information about all cohorts of passengers, a proportion of the lower risk passengers should also be screened in addition to all the higher risk passengers identified by statistical profiling (Robinson, Burgman, and R. Cannon, 2011). This extra inspection is for three reasons: i) it allows the monitoring of changes in risk for the lower predicted risk groups; ii) any update to the models used for

statistical profiling will require observations from the lower risk group in order to form unbiased predictions, and iii) it may act as a deterrent to non-targeted passengers.

In conclusion, we recommend that authorities investigate the use of targeted screening using statistical profiling for biosecurity risk material. However, larger, more comprehensive datasets should be assessed before implementation.

References

Breiman, L (2001). "Random forests". In: *Machine learning*. ISSN: 0885-6125.

Breiman, Leo et al. (2015). *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. R package version 4.6-12.

Burgman, Mark A (2016). *Trusting judgements: how to get the best out of experts*. Cambridge, United Kingdom Cambridge University Press, 2016.

Callegaro, Dario et al. (2016). "Development and external validation of two nomograms to predict overall survival and occurrence of distant metastases in adults after surgical resection of localised soft-tissue sarcomas of the extremities: a retrospective analysis". en. In: *The lancet oncology* 17.5, pp. 671–680. ISSN: 1470-2045, 1474-5488. DOI: 10.1016/S1470-2045(16)00010-3.

Cannon, R M (2009). "Inspecting and monitoring on a restricted budget–where best to look?" In: *Preventive veterinary medicine* 92.1-2, pp. 163–174. ISSN: 0167-5877, 1873-1716. DOI: 10.1016/j.prevetmed.2009.06.009.

Carpenter, Bob et al. (2016). "Stan: A probabilistic programming language". In: *Journal of Statistical Software (in press)*.

Crainiceanu, Ciprian M, David Ruppert, and Matthew P Wand (2005). "Bayesian Analysis for Penalized Spline Regression Using WinBUGS". In: *Journal of statistical software* 14.14, pp. 1–24. ISSN: 1548-7660.

Department of Agriculture and Water Resources (2016). *Annual Report 2015-16*. Tech. rep.

— (n.d.). *About our Biosecurity System*. http://www.agriculture.gov.au/biosecurity/australia/about. Accessed: 2016-3-NA.

Friedman, J (2001). "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics*. ISSN: 0090-5364.

Gelman, Andrew and Jennifer Hill (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press. ISBN: 9780511790942. DOI: 10.1017/CBO9780511790942.

Guo, Jiqiang, Jonah Gabry, and Ben Goodrich (2016). *rstan: R Interface to Stan*. R package version 2.9.0-3.

Hanley, J A and B J McNeil (1982). "The meaning and use of the area under a receiver operating characteristic (ROC) curve". en. In: *Radiology* 143.1, pp. 29–36. ISSN: 0033-8419. DOI: 10.1148/radiology.143.1.7063747.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York. ISBN: 9780387848570. DOI: 10.1007/978-0-387-84858-7.

Hua, Z, S Li, and Z Tao (2005). "A rule-based risk decision-making approach and its application in China's customs inspection decision". In: *The Journal of the Operational Research Society* 57.11, pp. 1313–1322. ISSN: 0160-5682, 1476-9360. DOI: 10.1057/palgrave.jors.2602142.

Hulme, Philip E (2009). "Trade, transport and trouble: managing invasive species pathways in an era of globalization". In: *The Journal of Applied Ecology* 46.1, pp. 10–18. ISSN: 0021-8901, 1365-2664. DOI: 10.1111/j.1365-2664.2008.01600.x.

IUCN (n.d.). *Species*. http://www.iucn.org/. Accessed: 2016-6-NA.

Kuhn, Max et al. (2016). *caret: Classification and Regression Training*. R package version 6.0-64.

Lai, Jyh-Mirn, Yi-Ting Hwang, and Chin-Cheng Chou (2012). "Modeling exotic highly pathogenic avian influenza virus entrance risk through air passenger violations". en. In: *Risk analysis: an official publication of the Society for Risk Analysis* 32.6, pp. 1093–1103. ISSN: 0272-4332, 1539-6924. DOI: 10.1111/j.1539-6924.2011.01740.x.

Leung, Brian et al. (2002). "An ounce of prevention or a pound of cure: bioeconomic risk analysis of invasive species". en. In: *Proceedings. Biological sciences / The Royal Society* 269.1508, pp. 2407–2413. ISSN: 0962-8452. DOI: 10.1098/rspb.2002.2179.

Lin, Xiao-Wei et al. (2009). "Foot-and-Mouth Disease Entrance Assessment Model Through Air Passenger Violations". In: *Risk analysis: an official publication of the Society for Risk Analysis* 29.4, pp. 601–611. ISSN: 0272-4332, 1539-6924. DOI: 10.1111/j.1539-6924.2008.01183.x.

Melo, Cristiano Barros de et al. (2014). "Profile of international air passengers intercepted with illegal animal products in baggage at Guarulhos and Galeão airports in Brazil". en. In: *SpringerPlus* 3, p. 69. ISSN: 2193-1801. DOI: 10.1186/2193-1801-3-69.

Park, Trevor and George Casella (2008). "The Bayesian Lasso". In: *Journal of the American Statistical Association* 103.482, pp. 681–686. ISSN: 0162-1459. DOI: 10.1198/016214508000000337. eprint: http://dx.doi.org/10.1198/016214508000000337.

Peltola, Tomi et al. (2014). "Hierarchical Bayesian Survival Analysis and Projective Covariate Selection in Cardiovascular Event Risk Prediction". In: *BMA@ UAI*, pp. 79–88.

Pimentel, David (2011). *Biological Invasions: Economic and Environmental Costs of Alien Plant, Animal, and Microbe Species*. Second. Hoboken: CRC Press, 2011.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ridgeway, Greg (2015). *gbm: Generalized Boosted Regression Models*. R package version 2.1.1.

Ripley, B D (2008). *Pattern Recognition and Neural Networks*. Cambridge University Press. ISBN: 9780521717700.

Ripley, Brian (2016). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. R package version 7.3-12.

Robinson, Andrew, Mark A Burgman, and Rob Cannon (2011). "Allocating surveillance resources to reduce ecological invasions: maximizing detections and information about the threat". In: *Ecological applications: a publication of the Ecological Society of America* 21.4, pp. 1410–1417. ISSN: 1051-0761.

Shih, Tai-Hwa, Chin-Cheng Chou, and Randall S Morley (2005). "Monte Carlo simulation of animal-product violations incurred by air passengers at an international airport in Taiwan". en. In: *Preventive veterinary medicine* 68.2-4, pp. 115–122. ISSN: 0167-5877. DOI: 10.1016/j.prevetmed.2004.11.010.

Sinden, J A et al. (2005). "The Economic Impact of Weeds in Australia". In: *Plant protection quarterly* 20.1, pp. 25–32. ISSN: 0815-2195. DOI: pes:2594.

Vilà, Montserrat et al. (2011). "Ecological impacts of invasive alien plants: a meta-analysis of their effects on species, communities and ecosystems". en. In: *Ecology Letters* 14.7, pp. 702–708. ISSN: 1461-023X, 1461-0248. DOI: 10.1111/j.1461-0248.2011.01628.x.

Winkler, Robert L (1967). "The Quantification of Judgment: Some Methodological Suggestions". In: *Journal of the American Statistical Association* 62.320, pp. 1105–1120. ISSN: 0162-1459. DOI: 10.2307/2283764.

Wood, S N (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

Wood, Simon N (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models". In: *Journal of the Royal Statistical Society. Series B, Statistical methodology* 73.1, pp. 3–36. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/j.1467-9868.2010.00749.x.

A Appendices

A.1 Model descriptions

Table 2 provides a full description of the models and the methods used for selecting tuning parameters (where required). All models were fit in `R` (R Core Team, 2016), with package `caret` (Kuhn et al., 2016) used to select tuning parameters for the GBM (Ridgeway, 2015), neural network (B. Ripley, 2016) and random forest (Leo Breiman et al., 2015) models. Tuning parameters for each prediction model were chosen by maximising predictive log-loss through repeated tenfold cross-validation, performed with five repetitions. Package `rstan` (Guo, Gabry, and Goodrich, 2016) was used to fit the Bayesian regression models via the MCMC sampler `stan` (Carpenter et al., 2016). The logistic regression model was fit using package `mgcv` (S N Wood, 2006). As discussed in Section 3.1, two separate sub-models were fit, based on passenger traits that do not need the IPC, and those that do. Table 2 provides detail for the second stage of models; the first stage is identical, but with a limited set of passenger traits used for modelling.

Table 2: Description of prediction models used for model comparison. If not specified, tuning parameters for each prediction model are chosen by maximising (predictive) log-loss through repeated tenfold cross-validation, performed with five repetitions. Appendix A.2 provides a brief overview of the tuning parameters for each method.

GAM	A logistic regression model. Occupation, visit reason and citizen group enter the model as categorical variables (see Table 1), whilst passenger age is modelled using thin-plate splines, with smoothing parameter selected via generalised cross-validation as implemented in the `mgcv` package. No interactions are included, and no variable selection or other model tuning is performed.
RF-caret	A random forest model with tuning parameter mtry – the number of randomly sampled split candidates $\in$ (2, 12, 23).
GBM-custom	A GBM with custom feature selection. Firstly, categorical variables are converted to dummy variables. Repeated subset sampling is then used to split the data into five training/testing datasets with a 70:30 split. For each (dummy) variable, a GBM is fit with tuning parameters set to: learning rate = 0.005; number of trees = 200; interaction depth = 2; and weights = 7 or 1 for observations that are non-compliant or compliant respectively. AUC is calculated for each individual model fit, and any variable with AUC < 0.51 is removed from further consideration. A GBM is fit to the remaining variables, and any variable with relative influence = 0 is removed. The **GBM-custom** model is a GBM model on the remaining variables.
GBM-caret	A GBM model with tuning parameters optimised over a grid of: learning rate $\in$ (0.1, 0.01, 0.005); number of trees $\in$ (700, 850, 1000); and interaction depth $\in$ (1, 2, 3).
NN-caret	A (single hidden layer) neural network model with tuning parameters over a grid of: size/number of units in hidden layer $\in$ (1, 3, 5); and weight decay $\in$ (0, 0.1, 0.0001).
Bayes-normal	Bayesian (shrinkage) logistic regression with normal priors on the regression coefficients. We assume a binomial model: $\Pr\left(Y_i = 1 \middle\vert X_i\right) = \text{logit}^{-1}\left(\mu + \beta_m M_i + \beta_d D_i + f(A_i) + \alpha_{o[i]} + \alpha_{v[i]} + \alpha_{c[i]}\right)$ $f(A_i) = \beta_a A_i + \sum_{k=1}^{K} \xi_k z_{ik}$ $\mu \sim N(0, 10)$ $\beta_j \sim N(0, \sigma_s^2)$, for $j = m, d, a$ $\xi_k \sim N(0, \sigma_a^2)$ $\alpha_{k[i]} \sim N(0, \sigma_s^2)$, for $k = o, v, c$ $\sigma_s \sim \text{half-}t_1(0, 1.0)$ $\sigma_a \sim \text{half-}t_1(0, 2.5)$

	where m, d denote male passengers and passengers who declared they had BRM respectively, a denotes passenger age, and o, v, c denote the categorical variables of passenger occupation, reason for visiting and citizenship group respectively (see Table 1). $f(\cdot)$ is a thin-plate spline, where the parameters ξ_k control the amount of shrinkage to the linear term (construction of the z_{ik} follows Crainiceanu, Ruppert, and Wand, 2005). Here the notation $\alpha_{k[i]}$ represents the coefficients for the *varying levels* of the k^{th} categorical variable (following Gelman and Hill, 2006).
Bayes-lasso	Bayesian (shrinkage) logistic regression with Laplace priors (Bayesian lasso model, e.g. Park and Casella, 2008; Peltola et al., 2014). The binomial model is the same as for the **Bayes-normal** model, with the following changes to the coefficient priors: $\beta_j \sim N(0, \sigma_s^2 \sigma_j^2)$, for $j = m, d$ $\alpha_{k[i]} \sim N(0, \sigma_s^2 \sigma_k^2)$, for $k = o, v, c, a$ $\sigma_s \sim$ half-$t_1(0, 1.0)$ $\sigma_j^2 \sim \text{Exp}(1)$, for $j = m, d$ $\sigma_k^2 \sim \text{Exp}(1)$, for $k = o, v, c, a$

A.2 Tuning parameters

In this appendix we provide some detail on the tuning parameters that are available in the prediction methods of Section 3. Hastie, Tibshirani, and Jerome Friedman (2009) provides an excellent overview of the methods and details of the tuning parameters.

Gradient boosting machine	
Number of trees	Controls the number of weak learners in the model. More trees lead to reduced error on the training set, but result in overfitting.
Learning rate, $\nu \in (0, 1]$	Provides regularisation by scaling the contribution of each weak learner to the current model fit. Lower values provide greater generalisation at a computational cost.
Interaction depth	Controls the maximum degree of interactions. A value of 2, for example, allows two-way interactions.
Neural network	
Weight decay	Controls the amount of shrinkage of the unknown weights towards 0. A weight decay value of 0 implies no shrinkage.
Size/number of units in hidden layer	The number of *derived features* in the hidden layer. These act similar to a basis expansion of the original predictors.
Random forest	
Number of randomly sampled split candidates	Number of predictor variables that are randomly sampled at each split of the weak learner.