

CEBRA Report Cover Page				
<b>Title, ID, &amp; Output #</b>	<i>Data Mining, CEBRA 1301A, Deliverable 3</i>			
<b>Project Type</b>	<i>Data Mining: Report for all Case Studies</i>			
<b>DAWR Project Sponsor</b>	<i>Raelene Vivian (was Tim Chapman)</i>	<b>DAWR Project Leader/s</b>	<i>Greg Hood</i>	
<b>CEBRA Project Leader</b>	<i>Andrew Robinson</i>	<b>NZ MPI Collaborator</b>	<i>Christine Reed</i>	
<b>Project Objectives</b>	<p><i>The Data Mining project will develop systems and protocols to inspect large volumes of biosecurity data. Broadly, the outcomes will improve the effectiveness and efficiency with which incoming cargo, mail, people, and vessels are screened into cohorts having different biosecurity risks—thereby ensuring that biosecurity risks are managed while minimising disruption of trade that meets the compliance standards of border agencies.</i></p> <p><i>Specifically, the project will coordinate a number of case studies that focus on different aspects of data mining. The coverage includes: (1) integration of biosecurity and geographical data; (2) using real-time passenger data from DIBP; (3) developing and assessing novel statistical patterns solely within existing data holdings to detecting anomalous broker activity; (4) risk factors in the Vessel Management System; (5) estimating risks and other values when there is scarce or incomplete evidence (transfer learning); (6) matching data mining outcomes and corporate statistics to give a 'broad brush' view of compliance across certain cargo pathways; (7) using hitchhiker pest location and activity information.</i></p>			
<b>Outputs</b>	<i>Three reports.</i>			
<b>CEBRA Workplan Budget</b>	<b>Year 2013-14</b>	<b>Year 2014-15</b>	<b>Year 2015-16</b>	<b>Year 2016-17</b>
	<i>\$240,000</i>	<i>\$78,000</i>	<i>TBA \$0,000</i>	<i>TBA \$0,000</i>
<b>Project Changes</b>	<p><i>The (Border) Compliance Division underwent substantial restructuring, including offering and acceptance of voluntary redundancies. This activity reduced the availability of department personnel for the sub-projects and the overall project from January 2014 onwards. It is likely that this restructuring and the redundancies had ongoing effects on the timeliness of the project delivery.</i></p> <p><i>Two of the seven sub-projects were terminated, namely sub-projects 5 and 7. Sub-project 5 was cancelled in order to allow more time and resources to focus upon several sub-projects that showed promise but needed follow-up analysis, for example, projects 1 and 2. Sub-project 7 was cancelled because the department's interception data were inadequate to complete the project.</i></p>			
<b>Research Outcomes</b>	<p><i>In each sub-project the tools were able to develop statistically reliable models that produced operationally realistic predictions. Specific summaries follow.</i></p> <ol style="list-style-type: none"> <li><i>1. Spatial Analysis of International Mail Interceptions — statistical analysis and maps of seizure data by census area are provided.</i></li> <li><i>2. Generalised Pattern Analysis for International Passengers (using real-time passenger data from DIBP) — data mining is complete and risk factors have been identified. Shortcomings are identified.</i></li> <li><i>3. Detecting anomalous broker activity — some suggestive patterns have been uncovered.</i></li> <li><i>4. Risk factors in the Vessel Management System — potential risk factors have been identified with many fields having some predictive power to detect failure. Shortcomings are identified.</i></li> <li><i>5. Transfer learning —terminated with agreement of the project sponsor.</i></li> <li><i>6. Developing a 'broad brush' view of compliance across certain cargo pathways (performance Indicators for CCV) — indicators of performance are presented and computed. Shortcomings are identified.</i></li> <li><i>7. Predicting hitchhiker activity —terminated with agreement of the project sponsor.</i></li> </ol>			
<b>Recommendations</b>	<p><i>17 recommendations are listed on page iii of the report:</i></p> <p><b><i>Spatial Analysis of International Mail Interceptions:</i></b></p> <ol style="list-style-type: none"> <li><i>1. Targeted public relations campaigns may reduce the incidence of biosecurity risk material being sent in the mail. Foreign-language university student associations should be asked to disseminate relevant biosecurity information to their members, and to assist in identifying legal sources of regulated goods that are commonly intercepted in the mail or air cargo pathways. A before-after control-intervention design should be used to assess evidence for the conjecture.</i></li> <li><i>2. Regulated goods are transported by air cargo as well as by international mail pathway. Spatial analysis similar to the mail analysis should be applied to destination addresses of seized articles in the air cargo pathway. Air cargo addresses are captured electronically, so the analysis will be more straightforward, and profiling of geographical hotspots for intervention would be possible.</i></li> </ol>			
<b>CEBRA Use only</b>	<b>Received by:</b>		<b>Date:</b>	
	CEBRA / SRC Approval		Date:	
	DAFF Endorsement ( ) Yes ( ) No		Date:	
	Report published		Date:	

	<p>3. The Sydney Gateway Facility has equipment that counts the international mail articles delivered by postcode. For Sydney data, obtain post-code specific delivery rates to estimate the relative risk of deliveries to specific locations and determine whether any are sufficiently risky to justify post-code targeting, as is done by the Australian Department of Immigration and Border Protection.</p> <p>4. The material risk of contamination transported by international mail may depend on the nature of its destination. Use the post-code of the delivery address to distinguish between urban and rural locations, and assess any effect upon the risk of contamination, from the points of view of (i) approach rate and (ii) magnitude of consequences.</p> <p>5. The declaration field on the international Customs form for international mail articles is hand-coded by the sender or their representative. Various factors cast doubt on the value of this information. Assess the veracity and utility of the international mail declaration field on the Customs form.</p> <p><b>Generalised Pattern Analysis for International Passengers:</b></p> <p>6. The potential benefits and challenges of international passenger screening using Australian Department of Immigration and Border Protection data should be assessed in terms of screening decision timeliness. Specifically, (i) will Border be willing and able to routinely make passenger data available to the department? and, (ii) will there be some means of flagging passengers for quarantine intervention at the primary line?</p> <p>7. Conditional on a positive outcome for Recommendation 6, develop a representative sample of passenger records for which the screened intervention is known and develop analytical techniques to better handle the different interception rates associated with the different intervention types.</p> <p>8. Conditional on a positive outcome for Recommendation 7, develop a snapshot survey using profiles developed from the screening model fitted to Immigration and Border Protection data, and inspect passenger cohorts rated as higher risk by either the Border data profiles or the departmental profiles. Analyse and compare the effectiveness of the profiles.</p> <p><b>Detecting anomalous broker activity:</b></p> <p>9. The department should note that there is some evidence in the broker analysis case study to suggest that brokers are altering CP status within declarations to reduce the apparent biosecurity risk after receiving a quarantine direction. Further analysis of this phenomenon may yield actionable intelligence.</p> <p>10. The broker analysis project did not involve the physical inspection of consignments for which tariff codes had been changed, which represents a possible avoidance behaviour by brokers. The department should consider expanding cargo surveillance to include those AIMS entries that are modified to appear less risky by the broker upon receipt of any quarantine directions.</p> <p>11. The department should note that the broker analysis project did not involve examination of goods that were not referred to the department because of provision of a low-risk answer to community protection profile questions. There may be some benefit to targeting such records as part of the Cargo Compliance Verification exercise (see Chapter 7).</p> <p><b>Risk factors in the Vessel Management System:</b></p> <p>12. For marine vessels, the identity of the agent was considered a priori to be a field that might best identify vessels with better or worse governance, and therefore possibly better or worse biosecurity compliance. The department should improve the quality of agent information for international marine vessels so that more inspection outcomes can be linked to the vessel agent, to provide a more rigorous test of this conjecture.</p> <p><b>Performance Indicators for CCV:</b></p> <p>13. CCV performance indicator interval estimates should be used by analysts to guide the temporal and hierarchical level of reporting of CCV performance indicator point estimates and to assist in the interpretation of the point estimates by pathway managers. The quality of the point estimates can be translated from the interval estimates to the pathway managers by the analysts (e.g., "the point estimate is poor").</p> <p>14. Performance indicators for CCV require information about the full range of goods imported to Australia. The department should obtain ICS data to enable computing system-wide measures of CCV performance indicators.</p> <p>15. CCV performance measures can be further distinguished by the policy area to which they are relevant by linking the profiles and tariffs to policy areas. The department should develop a code table that connects policy areas and tariffs to enable more fine-grained reporting for CCV performance measures.</p> <p>16. This project computed CCV performance measures for just two months, as a demonstration of</p>	
CEBRA Use only	Received by:	Date:
	CEBRA / SRC Approval	Date:
	DAFF Endorsement ( ) Yes ( ) No	Date:
	Report published	Date:



	<p>the underlying principles. The department should compute CCV performance measures for more time periods and consider reporting at a coarser scale, e.g. quarterly, or averaging across more than one month when reporting monthly.</p> <p><b>Hitch-hiker Interception Patterns:</b></p> <p>17. The hitch-hiker interception pattern analysis was significantly impeded by the lack of suitable interception data. Specific problems were inadequate linkages between the department's databases, and too few recorded instances of interception records for analysis. The department should develop more robust data capture and curation systems for gathering interception and operational data.</p>
Related Documents	Nil
Report Complete	Date: 6/5/2016

CEBRA Use only	Received by:	Date:
	CEBRA / SRC Approval	Date:
	DAFF Endorsement ( ) Yes ( ) No	Date:
	Report published	Date:

## 1301A Data Mining Final Report

---

CEBRA Project 1301A Deliverable 3

Sandy Clarke, Statistical Consulting Centre, The University of Melbourne

Andrew Robinson, CEBRA, The University of Melbourne

Matthew Chisholm, CEBRA, The University of Melbourne

Greg Hood, Border Compliance Division, Department of Agriculture and Water Resources

December 21, 2017

# Executive Summary

The Department of Agriculture and Water Resources (the department) seeks to mitigate the inherent biosecurity risk of various pathways by various control measures. This project involved the deployment of data-mining tools on a collection of data resources held by the department.

The overall results of the data mining exercises are very encouraging; in each completed sub-project the tools were able to develop statistically reliable models that produced operationally realistic predictions. Where appropriate, these models have been provided to the department for use in profiling. When the models led to insights that suggest further developments might be called for, these potential developments have been articulated. In the reports for all completed sub-projects, detailed recommendations have been made regarding further data that could be collected to improve modelling as well as the refinements to the modelling should such data be made available.

The List of Recommendations includes specific recommendations that arise from the case studies and corrective actions that may improve the accuracy and credibility of decisions arising from analyses of the data. Most of the recommended changes are procedural, carrying minimal if any additional cost.

Specific outcomes of the sub-projects are summarized below.

## 1. Spatial analysis of international mail interceptions.

- Aim: to explore the spatial distribution and patterns of the destination address of mail articles seized with biosecurity risk material.
- Outcome: spatial patterns of interceptions were identified, and linkages made with information from ABS census data. High-risk demographics were identified, and potential policy remedies discussed.
- Next: analysis could include incorporation of mail count data from the Sydney Gateway Facility, and land-use classification for each postcode. It is likely that the techniques used here will also be useful for profiling in the air cargo pathway.

## 2. Generalised pattern analysis for international passengers.

- Aim: to assess the value of Border passenger information for profiling international air passengers to improve inspection efficiency by (i) allowing more timely allocation of inspection resources, and (ii) more accurate profiles using more information than is available on the Incoming Passenger Card.
- Outcome: statistical models provide respectable predictive ability of the outcome of passenger inspection based on historical data, but the study suffers several critical shortcomings.
- Next: assess the potential benefits and challenges of implementing any outcomes from the study before trying to correct the shortcomings.

## 3. Detecting anomalous broker activity.

- Aim: to outline analyses of the propensity of brokers to amend declarations after receiving quarantine directions of different nature.
- Outcome: very little evidence of a link between alterations to import declarations and quarantine directions of different kinds among the brokers assessed in this analysis. Some evidence of alteration of CP risk status from high to low upon direction.
- Next: the department should note these results and consider surveillance on items that correspond with amended declarations.

## 4. Risk factor extraction with VMS.

- Aim: to identify risk factors using data from the Vessel Management System (VMS) dataset, in order to be able to predict inspection failure.
  - Outcome: the aim of determining risk factors for vessels has been achieved with the data available.
  - Next: improve data capture for vessel agents (75% missing in dataset) and consider future analysis along comparable lines.
5. Transfer Learning.
- Aim: to explore possibilities of sharing information among similar but arguably unlike pathways.
  - Outcome: this sub-project was withdrawn owing to time and resource constraints.
  - Next: no further steps at this time.
6. Performance indicators for Cargo Compliance Verification.
- Aim: to develop measures that allow reporting compliance across all cargo pathways.
  - Outcome: Performance measures are developed and reported.
  - Next: obtain Customs data of all cargo to enable the calculation of key system-wide measures. Produce summaries at the policy level.
7. Analysis of hitch-hikers interception data.
- Aim: to test the hypothesis that the arrival of specific hitchhiker pests on imported cargo can be anticipated based on pest biology and known distribution.
  - Outcome: this sub-project was withdrawn as insufficient interception data were available, and those data that were available were of inadequate quality.
  - Next: the department should identify and remedy the gaps in data capture and curation that undermined the successful prosecution of this study.

## Acknowledgments

The authors are grateful to the following Department of Agriculture and Water Resources personnel for their substantial contribution to this work; Jose Arias, Wayne Atkinson, Jamie Brown, Dr. Peta Holmes, Robert Kancans, Alan Küffer, Nianjun Liu, Claire McKee, Kathleen Quan, Stephen Richardson, Dr. Nyree Stenekes, Wayne Terpstra, and Chris Woodland.

# List of Recommendations

1	Targeted public relations campaigns may reduce the incidence of biosecurity risk material being sent in the mail. Foreign-language university student associations should be asked to disseminate relevant biosecurity information to their members, and to assist in identifying legal sources of regulated goods that are commonly intercepted in the mail or air cargo pathways. A before-after control-intervention design should be used to assess evidence for the conjecture. . . . .	14
2	Regulated goods are transported by air and sea cargo as well as by international mail. Spatial analysis similar to the mail analysis should be applied to destination addresses of seized articles in the air and sea cargo pathways. Cargo addresses are captured electronically, so the analysis will be more straightforward, and profiling of geographical hotspots for intervention would be possible. . . . .	14
3	The Sydney Gateway Facility has equipment that counts the international mail articles delivered by postcode. For Sydney data, obtain post-code specific delivery rates to estimate the relative risk of deliveries to specific locations and determine whether any are sufficiently risky to justify post-code targeting, as is done by the Australian Department of Immigration and Border Protection. . . . .	14
4	The material risk of contamination transported by international mail may depend on the nature of its destination. Use the post-code of the delivery address to distinguish between urban and rural locations, and assess any effect upon the risk of contamination, from the points of view of (i) approach rate and (ii) magnitude of consequences. . . . .	14
5	The declaration field on the international Customs form for international mail articles is hand-coded by the sender or their representative. Various factors cast doubt on the value of this information. Assess the veracity and utility of the international mail declaration field on the Customs form. . . . .	14
6	The potential benefits and challenges of international passenger screening using Australian Department of Immigration and Border Protection data should be assessed in terms of screening decision timeliness. Specifically, (i) will Border be willing and able to routinely make passenger data available to the department? and, (ii) will there be some means of flagging passengers for quarantine intervention at the primary line? . . . . .	22
7	Conditional on a positive outcome for Recommendation 6, develop a representative sample of passenger records for which the screened intervention is known and develop analytical techniques to better handle the different interception rates associated with the different intervention types. . . . .	23
8	Conditional on a positive outcome for Recommendation 7, develop a snapshot survey using profiles developed from the screening model fitted to Immigration and Border Protection data, and inspect passenger cohorts rated as higher risk by either the Border data profiles or the departmental profiles. Analyse and compare the effectiveness of the profiles. . . . .	23
9	The department should note that there is evidence in the broker analysis case study to suggest that some brokers are altering CP status within declarations to reduce the apparent biosecurity risk after receiving a quarantine direction. Further analysis of this phenomenon may provide actionable intelligence. . . . .	34
10	The broker analysis project did not involve the physical inspection of consignments for which tariff codes had been changed, which represents a possible avoidance behaviour by brokers. The department should consider expanding cargo surveillance to include those AIMS entries that are modified to appear <i>less risky</i> by the broker upon receipt of any quarantine directions. . . .	34
11	The department should note that the broker analysis project did not involve examination of goods that were not referred to the department because of provision of a low-risk answer to community protection profile questions. There may be some benefit to targeting such records as part of the Cargo Compliance Verification exercise (see Chapter 7) . . . . .	34

12	For marine vessels, the identity of the agent was considered <i>a priori</i> to be a field that might best identify vessels with better or worse governance, and therefore possibly better or worse biosecurity compliance. The department should improve the quality of agent information for international marine vessels so that more inspection outcomes can be linked to the vessel agent, to provide a more rigorous test of this conjecture. . . . .	40
13	CCV performance indicator interval estimates should be used by analysts to guide the temporal and hierarchical level of reporting of CCV performance indicator point estimates and to assist in the interpretation of the point estimates by pathway managers. The quality of the point estimates can be translated from the interval estimates to the pathway managers by the analysts (e.g., “the point estimate is poor”). . . . .	51
14	Performance indicators for CCV require information about the full range of goods imported to Australia. The department should obtain ICS data to enable computing system-wide measures of CCV performance indicators. . . . .	53
15	CCV performance measures can be further distinguished by the policy area to which they are relevant by linking the profiles and tariffs to policy areas. The department should develop a code table that connects policy areas and tariffs to enable more fine-grained reporting for CCV performance measures. . . . .	53
16	This project computed CCV performance measures for just two months, as a demonstration of the underlying principles. The department should compute CCV performance measures for more time periods and consider reporting at a coarser scale, e.g. quarterly, or averaging across more than one month when reporting monthly. . . . .	53
17	The hitch-hiker interception pattern analysis was significantly impeded by the lack of suitable interception data. Specific problems were inadequate linkages between the department’s databases, and too few recorded instances of interception records for analysis. The department should develop more robust data capture and curation systems for gathering interception and operational data. . . . .	55



# Contents

<b>Executive Summary</b>	<b>i</b>
<b>List of Recommendations</b>	<b>iii</b>
<b>Table of Contents</b>	<b>v</b>
<b>List of Tables</b>	<b>viii</b>
<b>List of Figures</b>	<b>ix</b>
<b>Table of Definitions</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Spatial Analysis of International Mail Interceptions</b>	<b>3</b>
2.1 Summary . . . . .	3
2.1.1 Background . . . . .	3
2.1.2 Motivating Question . . . . .	3
2.1.3 Methods . . . . .	3
2.1.4 Results . . . . .	3
2.1.5 Conclusions and Future Directions . . . . .	4
2.2 Data preparation . . . . .	4
2.3 Spatial mapping . . . . .	8
2.4 Relating seizure counts and population characteristics . . . . .	9
2.4.1 Analysis approach . . . . .	9
2.4.2 Description and interpretation of results . . . . .	11
2.4.3 Summary of key results . . . . .	12
2.4.4 Additional results . . . . .	13
2.5 Conclusion and Recommendations . . . . .	14
<b>3 Generalised Pattern Analysis for International Passengers</b>	<b>15</b>
3.1 Summary . . . . .	15
3.1.1 Background . . . . .	15
3.1.2 Motivating Question . . . . .	15
3.1.3 Methods . . . . .	15
3.1.4 Results . . . . .	16
3.1.5 Conclusions and Future Directions . . . . .	16
3.2 Data preparation . . . . .	16
3.3 Data analysis . . . . .	17
3.4 Results . . . . .	19
3.5 Implementation Guidelines . . . . .	21
3.6 Conclusion and Recommendations . . . . .	22
<b>4 Detecting Anomalous Broker Activity</b>	<b>24</b>
4.1 Summary . . . . .	24
4.1.1 Background . . . . .	24
4.1.2 Motivating Question . . . . .	24
4.1.3 Methods . . . . .	25
4.1.4 Results . . . . .	25

4.1.5	Conclusions and Future Directions . . . . .	25
4.2	Data preparation . . . . .	25
4.3	Data analysis . . . . .	25
4.3.1	The impact of threatening amendments . . . . .	26
4.3.2	The impact of special directions . . . . .	30
4.3.3	Community Protection questions and amendments . . . . .	33
4.4	Conclusion and Recommendations . . . . .	33
<b>5</b>	<b>Risk Factor Prediction for International Vessels</b>	<b>35</b>
5.1	Summary . . . . .	35
5.1.1	Background . . . . .	35
5.1.2	Motivating Question . . . . .	35
5.1.3	Methods . . . . .	35
5.1.4	Results . . . . .	35
5.1.5	Conclusions and Future Directions . . . . .	36
5.2	Data preparation . . . . .	36
5.3	Data analysis . . . . .	37
5.4	Results . . . . .	38
5.5	Conclusion and Recommendations . . . . .	40
<b>6</b>	<b>Overview of Transfer Learning</b>	<b>41</b>
6.1	Summary . . . . .	41
6.1.1	Background . . . . .	41
6.1.2	Motivating Question . . . . .	41
6.1.3	Methods . . . . .	41
6.1.4	Results . . . . .	41
6.1.5	Conclusions and Future Directions . . . . .	41
6.2	Relevance . . . . .	41
6.3	Terminology . . . . .	42
6.3.1	Formal transfer learning terminology . . . . .	42
6.4	How to transfer . . . . .	43
6.4.1	Instance transfer . . . . .	43
6.4.2	Feature representation transfer . . . . .	43
6.4.3	Parameter transfer . . . . .	43
6.4.4	Relational knowledge transfer . . . . .	44
6.4.5	Available resources . . . . .	44
6.5	When to transfer . . . . .	44
6.6	Examples . . . . .	44
6.6.1	Image recognition . . . . .	44
6.6.2	Border compliance . . . . .	45
6.7	Conclusion and Recommendations . . . . .	45
<b>7</b>	<b>Performance Indicators for Cargo Compliance Verification</b>	<b>46</b>
7.1	Summary . . . . .	46
7.1.1	Background . . . . .	46
7.1.2	Motivating Question . . . . .	46
7.1.3	Methods . . . . .	47
7.1.4	Results . . . . .	47
7.1.5	Conclusions and Future Directions . . . . .	47
7.2	Data preparation . . . . .	47
7.3	Monthly CCV performance . . . . .	47
7.4	Monthly system-wide measures . . . . .	47
7.5	Precision . . . . .	51
7.6	Conclusion and Recommendations . . . . .	53
<b>8</b>	<b>Interception Patterns of Hitch-hiker Pests</b>	<b>54</b>
8.1	Summary . . . . .	54
8.1.1	Background . . . . .	54
8.1.2	Motivating Question . . . . .	54

8.1.3	Methods . . . . .	54
8.1.4	Results . . . . .	54
8.1.5	Conclusions and Future Directions . . . . .	54
8.2	Datasets . . . . .	54
8.3	Organisms of interest . . . . .	55
8.4	Potential organisations of interest . . . . .	55
8.4.1	United Nations FAO . . . . .	55
8.4.2	IPPC . . . . .	56
8.4.3	Regional Plant Protection Organisations . . . . .	57
8.4.4	Secretariat of the Pacific Community . . . . .	58
8.4.5	Pacific Islands Development Program . . . . .	58
8.4.6	CIRAD . . . . .	59
8.4.7	Agrisles . . . . .	59
8.4.8	Organisation for Economic Cooperation and Development (OECD) . . . . .	59
8.4.9	European Union . . . . .	59
8.4.10	National organisations . . . . .	59
8.5	Sub-national organisations . . . . .	62
8.6	Search method . . . . .	62
8.7	Other projects . . . . .	63
<b>Literature Cited</b>		<b>63</b>
<b>A Random forests overview</b>		<b>66</b>
<b>B Spatial Analysis of International Mail Interceptions</b>		<b>68</b>
B.1	modellingSA2.R . . . . .	68
B.2	spatialplots.R . . . . .	77
<b>C Generalised Pattern Analysis for International Passengers</b>		<b>82</b>
<b>D Detecting Anomalous Broker Activity</b>		<b>84</b>
<b>E Risk Factor Extraction with VMS</b>		<b>94</b>
E.1	VMSscript.R . . . . .	94
E.2	vmsfunctions.R . . . . .	98
E.3	Performance Indicators for Cargo Compliance Verification . . . . .	100
E.3.1	Tables of results for July 2013 . . . . .	100

# List of Tables

2.1	Key seizure fields provided from MAPS data seizure records (S) and ABS data from geocoding (G) for spatial analysis of mail seizures. . . . .	5
2.2	Language distinctions provided in ABS census data for spatial analysis of mail seizures. .	6
2.3	Demographic combinations provided in ABS census data for spatial analysis of mail seizures.	7
2.4	Counts of seizure classes in the mail pathway 2008–2011, sorted by decreasing frequency. .	10
2.5	Counts of seizure types in the mail pathway by key commodity 2008–2011, sorted by descending frequency. . . . .	10
3.1	Comparing the frequencies of the top 10 travel document countries with their frequencies in the IPC sample. . . . .	17
3.2	Comparing the frequencies of the top 10 flight numbers with their frequencies in the IPC sample. . . . .	17
3.3	Data fields available for international passenger risk analysis. . . . .	18
3.4	Data fields for international passenger risk analysis after initial feature selection. . . . .	19
3.5	Relative importance and crude risk for international passenger data, for the top 10 field levels. . . . .	20
3.6	Performance of the full model on withheld passenger inspection data (with proportions). .	21
3.7	Predictions for simple passenger profile model. . . . .	22
3.8	Performance of simple predicted passenger profile model. . . . .	22
4.1	Broker behavior data: Amendment counts and rates per direction, ranked by amendment rate. . . . .	26
4.2	Broker record amendment odds ratios after threatening directions, with a 95% confidence interval (CI) and count of each amendment type. . . . .	27
4.3	Broker amendment odds ratios after special directions, with a 95% confidence interval (CI) and count of each amendment type. . . . .	30
4.4	Incidence of any high risk answer from the broker before and after amendment with row percentages. . . . .	33
4.5	Broker amendment odds ratio for CP high risk removal, with a 95% confidence interval (CI) and percentage of CP high risk reductions that include that type of amendment. . .	33
5.1	Potential factors currently available for modelling biosecurity risk in vessels. . . . .	36
5.2	Data fields for biosecurity risk analysis after initial feature selection in vessel inspection data. . . . .	38
5.3	Relative importance, crude risks, odds ratio and rarity, for top 18 field levels for modeling biosecurity risk from vessel inspection data. . . . .	39
5.4	Performance of the vessel inspection prediction model for biosecurity risk on test data, for a variety of sampling fractions. . . . .	39
7.1	CCV failure rates by country. . . . .	48
7.2	Definition of nodes including counts for July 2013. A dash indicates that the datum for this column is not relevant to the calculation, and ‘NA’ is an actual level of the variable considered. . . . .	50
7.3	95% confidence intervals for BIC and PIC, for entries in July 2013. . . . .	52
E.1	CCV failure rates by profile pathway . . . . .	100
E.2	CCV failure rates by country . . . . .	100
E.3	CCV failure rate by processing state . . . . .	100
E.4	CCV failure rate by broker . . . . .	101

E.5	CCV failure rate by importer code . . . . .	102
E.6	CCV failure rate by profile code . . . . .	103
E.7	CCV failure rate by supplier code . . . . .	106
E.8	CCV failure rate by tariff code . . . . .	107

# List of Figures

2.1	Greater Melbourne destinations to which seized international mail articles were addressed (raw points) with map. . . . .	8
2.2	Greater Melbourne destinations to which seized international mail articles were addressed (counts) with map. . . . .	9
2.3	Some example partial plots for the spatial analysis of mail interceptions, presenting the estimated overall relationship between two key demographic variables and predicted seizure count, based on the random forest model. . . . .	11
2.4	A partial plot for the top five languages for predicting total seizures within a census region in the mail pathway. . . . .	13
3.1	Distribution of international passenger inspection records over time. . . . .	16
3.2	The relationship between age and the predicted probability of non-compliance for international passengers. . . . .	20
4.1	Magnitude of changes by brokers in delivery address by threatening direction. . . . .	28
4.2	Magnitude of changes by broker of importer address by threatening direction. . . . .	28
4.3	Absolute difference in number of unique tariff classes by threatening direction. . . . .	28
4.4	Absolute difference in the number of lines by threatening direction. . . . .	29
4.5	Absolute difference in the number of unique quantities by threatening direction. . . . .	29
4.6	Absolute difference in the number of unique goods by threatening direction. . . . .	29
4.7	Magnitude of changes by brokers in delivery address by special direction. . . . .	31
4.8	Magnitude of changes in importer address by special direction. . . . .	31
4.9	Absolute difference in number of unique tariff classes by special direction. . . . .	31
4.10	Absolute difference in the number of lines by special direction. . . . .	32
4.11	Absolute difference in the number of unique quantities by special direction. . . . .	32
4.12	Absolute difference in the number of unique goods by special direction. . . . .	32
5.1	The performance of the vessel risk prediction model on test data, depicted as a ROC curve. . . . .	40
7.1	The cargo compliance verification (CCV) summary for July 2013, with more detailed failure outlines. . . . .	48
7.2	A bar chart demonstrating the failure rates for a range of tariff codes for July 2013 with lines (entries). . . . .	49
7.3	System-wide CCV flowchart with letters assigned to each node. . . . .	50
7.4	Confidence intervals for BIC under CCV, for entries in July. . . . .	52
7.5	Confidence intervals for PIC under CCV, for entries in July. . . . .	52
7.6	An example of a graphic designed to show changes over time, with estimate and 95% confidence interval based on CCV results. . . . .	53
A.1	An example of a tree. . . . .	66

## Table of Definitions

Term	Definition
ABARES	<i>Australian Bureau of Agricultural and Resource Economics and Sciences</i> is a research bureau within the Department of Agriculture and Water Resources.
ABS	<i>Australian Bureau of Statistics</i> .
AIMS	<i>Agriculture Information Management System</i> .
BIC	<i>Before Intervention Compliance</i> is the proportion of units that comply with biosecurity regulations before an intervention is performed.
Biosecurity risk material	<i>Biosecurity risk material</i> is any material perceived to pose a threat to Australia's food security, unique environment and economy, due to the likely presence of exotic pests and diseases.
CCV	<i>Cargo Compliance Verification</i> was, at the time of writing, a system of inspections conducted on containerised sea cargo to assess the integrity of the Australian biosecurity system.
Categorical field	A <i>categorical field</i> is a variable which takes on a finite number of discrete levels.
Demographic field	A <i>demographic field</i> is a measurable quantity for a given population or sub-population at a given point of time.
Feature selection	<i>Feature selection</i> is the initial process of reducing the number of fields considered.
Field	A <i>field</i> or variable is a measurable feature that can take different values.
Geocoding	<i>Geocoding</i> is the process of converting a description of a location (e.g. a postal address) into geographical coordinates (e.g. latitude and longitude).
MAPS	<i>Mail and Passenger System</i> is a departmental database that records the regulatory effort and outcomes for international passengers and mail.
Non-compliant	For our purposes, <i>non-compliance</i> refers to the occurrence of undeclared biosecurity risk material in passengers, mail, or cargo.
Numerical field	A <i>numerical field</i> is a measured field which take a numerical or quantitative value.
Outcome	An <i>outcome</i> or response is the event of interest that we are seeking to predict.
Overfitting	In statistics, <i>overfitting</i> is when a model captures random error instead of the underlying relationship, because there are too many predictors relative to the number of observations.
PDC	<i>Pratique Documentary Clearance</i> is a risk-based algorithm used by the department to reduce intervention effort systematically for low-risk vessels with a proven history of compliance.
PIC	<i>Post-Intervention Compliance</i> is the proportion of units that comply with biosecurity regulations after an intervention is performed.
Predictors	<i>Predictors</i> or explanatory fields are a set of characteristics of an observation or unit that may be used to predict an outcome.
ROC curve	A <i>receiver operating characteristic curve</i> is a graphical plot that illustrates the performance of a binary classifier, by plotting the true positive rate against the false positive rate for a range of thresholds of the classifier.
Sparse	A <i>sparse</i> field (or level of a field) is one that is infrequently present, commonly taking the value of zero.
SA1	<i>Statistical Area 1</i> is the smallest spatial unit for the release of ABS Census data.
SA2	<i>Statistical Area 2</i> is a medium spatial unit for the release of ABS Census data.
VMS	<i>Vessel Management System</i> is the database used by the department for managing the biosecurity risk of international ocean-going vessels.

# Chapter 1

## Introduction

This report provides a summary of the final results of the sub-projects in this data mining project, namely:

1. Spatial analysis international mail interceptions;
2. Generalised pattern analysis for international passengers;
3. Detecting anomalous broker activity;
4. Risk factor extraction with VMS;
5. Transfer learning;
6. Performance indicators for Cargo Compliance Verification; and
7. Analysis of hitch-hikers interception data.

Each of these sub-projects tried to use the wealth of data available to the department for the purposes of exploration and future planning. The methodological focus was data mining, which is a suite of modern statistical techniques that are designed to gain useful information from large data sets.

In particular, we used *random forests* modelling, which is a flexible, powerful approach to constructing models for which the key application is prediction, as opposed to the estimation of parameters of interest. An overview of this model-fitting strategy is provided in Appendix A. Random forests were used for Chapter 2. We also used gradient boosting machines, which are related to random forests but reweight the observations by how poor the previous fits were for them, in Chapters 3 and 5. Finally we used logistic regression for the analysis reported in Chapter 4, because the data were entirely categorical.

Data mining methods such as random forests modeling have several important advantages over classical statistical techniques, such as linear and generalized linear modeling. These advantages include, but are not limited to:

- Handling large data and sparse sets. Computationally, these methods can handle a very large number of input records and will produce models for datasets for which traditional methods will fail.
- Modelling non-linear relationships. These methods are highly flexible, and require no assumptions about the underlying nature of the relationship between predictors and outcomes (e.g., that the true nature of the relationship is linear, or even continuous).
- Modelling complex relationships between predictors. Interactions between predictors can be very important in the prediction of outcomes, and can be easily incorporated into these methods.
- Avoiding overfitting. These methods automatically involve repeated modelling on random subsets of data, to ensure the chosen model performs well on new data, avoiding spurious results.
- Powerful results. These methods have been shown to have greater predictive power compared with classical techniques. For a demonstration of the performance advantages of random forests models in particular, see Fernández-Delgado et al. (2014).

Furthermore, the modern model-fitting techniques offer comparably easy implementation in software. These methods can be easily implemented in statistical software such as **R** (R Core Team, 2013), the software used in this project. An excellent reference for more technical detail on data mining methods is Hastie et al. (2009).

These approaches typically involve a trade-off between predictive power and interpretative power, but we have sought to develop high-power predictors alongside summaries to aid interpretation for each sub-project. The chapters for each sub-project describe the specific methodology used and provide specific advice for interpretation, relating to the aims of the sub-project.



## Chapter 2

# Spatial Analysis of International Mail Interceptions

SANDY CLARKE\*, NYREE STENEKES<sup>†</sup>, ROBERT KANCANS<sup>†</sup>, CHRIS WOODLAND<sup>‡</sup>, AND ANDREW ROBINSON<sup>§</sup>

### 2.1 Summary

#### 2.1.1 Background

This chapter<sup>1</sup> focuses on the characteristics of international mail articles that have been intercepted with biosecurity risk material (BRM) after inspection at one of the Gateway Facilities. The biosecurity risk of select international mail articles is routinely assessed upon arrival by x-ray or detector dog units. If BRM is detected then the article is seized, and its details recorded.

#### 2.1.2 Motivating Question

What are the spatial patterns of the delivery addresses of seized international mail articles, and how do these relate to demographic features such as language, education, employment and age?

#### 2.1.3 Methods

We used all seizures from 1 July 2002 until 26 June 2012 in the Mail and Passenger System (MAPS) database and demographic fields based on 2011 Australian Bureau of Statistics census data at the levels of statistical area 1 (SA1) and statistical area 2 (SA2). We cleaned and geo-located the recorded addresses automatically and then applied spatial statistical tools as well as fitting a random forest model.

#### 2.1.4 Results

The results of this analysis confirm some existing opinions regarding common demographics of destinations with high seizure counts, particularly that higher seizure counts often occur in areas with large numbers of young Chinese speakers, such as those attending university. However, there are some other, more subtle relationships with individual commodities such as a correlation between live plants and non-English European languages. These results provide a rich picture of the composition of high seizure environments, enabling specifically targeted campaigns.

---

\*Statistical Consulting Centre, The University of Melbourne

<sup>†</sup>Australian Bureau of Agricultural and Resource Economics and Sciences, Department of Agriculture and Water Resources

<sup>‡</sup>Compliance Division, Department of Agriculture and Water Resources

<sup>§</sup>CEBRA, The University of Melbourne

<sup>1</sup>This chapter is here reproduced almost verbatim from Clarke et al. (2015a), for completeness of this report. The research work performed for this chapter, with sensitive details omitted, has been presented as Clarke et al. (2015b). Recommendations and the Summary section have been added.

### 2.1.5 Conclusions and Future Directions

The methods can be used for future seizure data in mail, and could also be applied to sea and air cargo inspection data, with the advantage that delivery details could be used to construct profiles, as delivery address is captured electronically for these pathways.

## 2.2 Data preparation

Two sets of data were made available to analyse the relationship between seizure locations and population characteristics or demographics of the local area. First, the Mail and Passenger System (MAPS) database provided all seizures from 1 July 2002 until 26 June 2012, including a variety of characteristics of the seizure, the most relevant of which were the addresses and the nature of the seized goods. The list of available fields is given in Table 2.1, including both those fields from the original departmental seizures and the information generated by the geocoder based on the destination address.

The second set of data, provided by ABARES, contained the demographic fields, which were based on ABS data at the levels of statistical area 1 (SA1) and statistical area 2 (SA2) from 2011. SA1s are the smallest unit for the release of Census data, which contains approximately 55,000 SA1 spatial units. SA2s are medium-sized aggregates of SA1s; the census contains approximately 2,196 SA2 spatial units. These fields were deliberately chosen as likely to be related to seizure counts as well as being potential targets for public awareness campaigns, and include languages, educational status, age and certain agricultural employment levels. Counts of combinations of these fields were also provided.

Table 2.2 gives a list of all the languages provided and Table 2.3 gives the sub-categorizations of Chinese speakers provided; a dash here means the count was across the full range of levels for that demographic field in that instance.

Initially the demographics were provided at the SA1 level, but this was too high resolution for analysis, because the small units contained only small counts of seizures and comparatively sparse demographic data. Subsequently, demographic data were provided at the SA2 level. SA2 codes can be generated directly from SA1 codes. All seizures with addresses that could be matched by the geocoding software could then be assigned an SA2 code. Seizures with no address match or a tie for an address match were excluded, as well as seizures which were declared appropriately (and therefore not a quarantine risk). Counts of these seizures (both totals and specific categories or commodities) could then be aggregated at the SA2 level, and matched to relevant demographics via statistical modeling.

Note that the counts of seizures of various types aggregated at the SA2 level have been provided as both totals and counts per 100,000 persons in the SA2 area for identification of key SA2 areas.

**Table 2.1:** Key seizure fields provided from MAPS data seizure records (S) and ABS data from geocoding (G) for spatial analysis of mail seizures.

Seizure field	Source	Description
Loc_name	G	GNAF, STLOC, or XROADS
Status	G	Match, mismatch or tie
Score	G	Match score
Match_address	G	Address matched to
Ref_ID	G	Address ID
X	G	Longitude
Y	G	Latitude
GC_Flag	G	0, A, L or S
LGA_Name	G	ABS local government area
Meshblock	G	ABS spatial unit
SA1_2011	G	Statistical Area 1
ARC_Street	G	
ARC_City	G	
ARC_State	G	
ARC_ZIP	G	
Field1	G	Geocoder ID
seizureid	S	Seizure ID
inspection	S	Date of inspection
category	S	Category of seizure
subcategory	S	Subcategory of seizure
commodity	S	Commodity seized
countryoforigin	S	
declaration	S	Nature of declaration
surnamecompanyname	S	
firstname	S	
address	S	
citytownsuburb	S	
state	S	
postcode	S	
country	S	
phonenumbers	S	
statuslocation	S	Location of seizure
persontype	S	Consignee, payee or sender address

**Table 2.2:** Language distinctions provided in ABS census data for spatial analysis of mail seizures.

Northern European	Hmong Mien
Celtic	Mon Khmer
English	Tai
German Related	Southeast Asian Austronesian
Dutch Related	Other Southeast Asian
Scandinavian	Eastern Asian
Finnish Related	Chinese
Southern European	Japanese
French	Korean
Greek	Other Eastern Asian
Iberian	Australian Indigenous
Italian	Arnhem Land Daly River Region
Maltese	Yolngu Matha
Other Southern European	Cape York Peninsula
Eastern European	Torres Strait Island
Baltic	Northern Desert Fringe Area
Hungarian	Arandic
East Slavic	Western Desert
South Slavic	Kimberley Area
West Slavic	Other Australian Indigenous
Other Eastern European	Other
Southwest and Central Asian	American
Iranic	African
Middle Eastern Semitic	Pacific Austronesian
Turkic	Oceanian Pidgins Creoles
Other Southwest Central Asian	Papua New Guinea
Southern Asian	Invented
Dravidian	Sign
Indo Aryan	Supplementary codes
Other Southern Asian	Not stated
Southeast Asian	Total
Burmese Related	

**Table 2.3:** Demographic combinations provided in ABS census data for spatial analysis of mail seizures.

Age	Language	Employment	Education
-	Chinese	Nursery Floriculture	-
-	Chinese	Mushroom Veg	-
-	Chinese	Fruit Nut	-
-	Chinese	Sheep Beef	-
-	Chinese	Dairy Cattle	-
-	Chinese	Poultry	-
-	Chinese	Deer	-
-	Chinese	Other Livestock	-
0-14	Chinese	-	Technical Further
15-24	Chinese	-	Technical Further
25-39	Chinese	-	Technical Further
40-64	Chinese	-	Technical Further
65+	Chinese	-	Technical Further
-	Chinese	-	Technical Further
0-14	Chinese	-	University
15-24	Chinese	-	University
25-39	Chinese	-	University
40-64	Chinese	-	University
65+	Chinese	-	University
-	Chinese	-	University
0-14	Chinese	-	Other
15-24	Chinese	-	Other
25-39	Chinese	-	Other
40-64	Chinese	-	Other
65+	Chinese	-	Other
-	Chinese	-	Other
0-14	Chinese	-	Not stated
15-24	Chinese	-	Not stated
25-39	Chinese	-	Not stated
40-64	Chinese	-	Not stated
65+	Chinese	-	Not stated
-	Chinese	-	Not stated
0-14	Chinese	-	Not applicable
15-24	Chinese	-	Not applicable
25-39	Chinese	-	Not applicable
40-64	Chinese	-	Not applicable
65+	Chinese	-	Not applicable
-	Chinese	-	Not applicable
0-14	-	-	-
15-24	-	-	-
25-39	-	-	-
40-64	-	-	-
65+	-	-	-

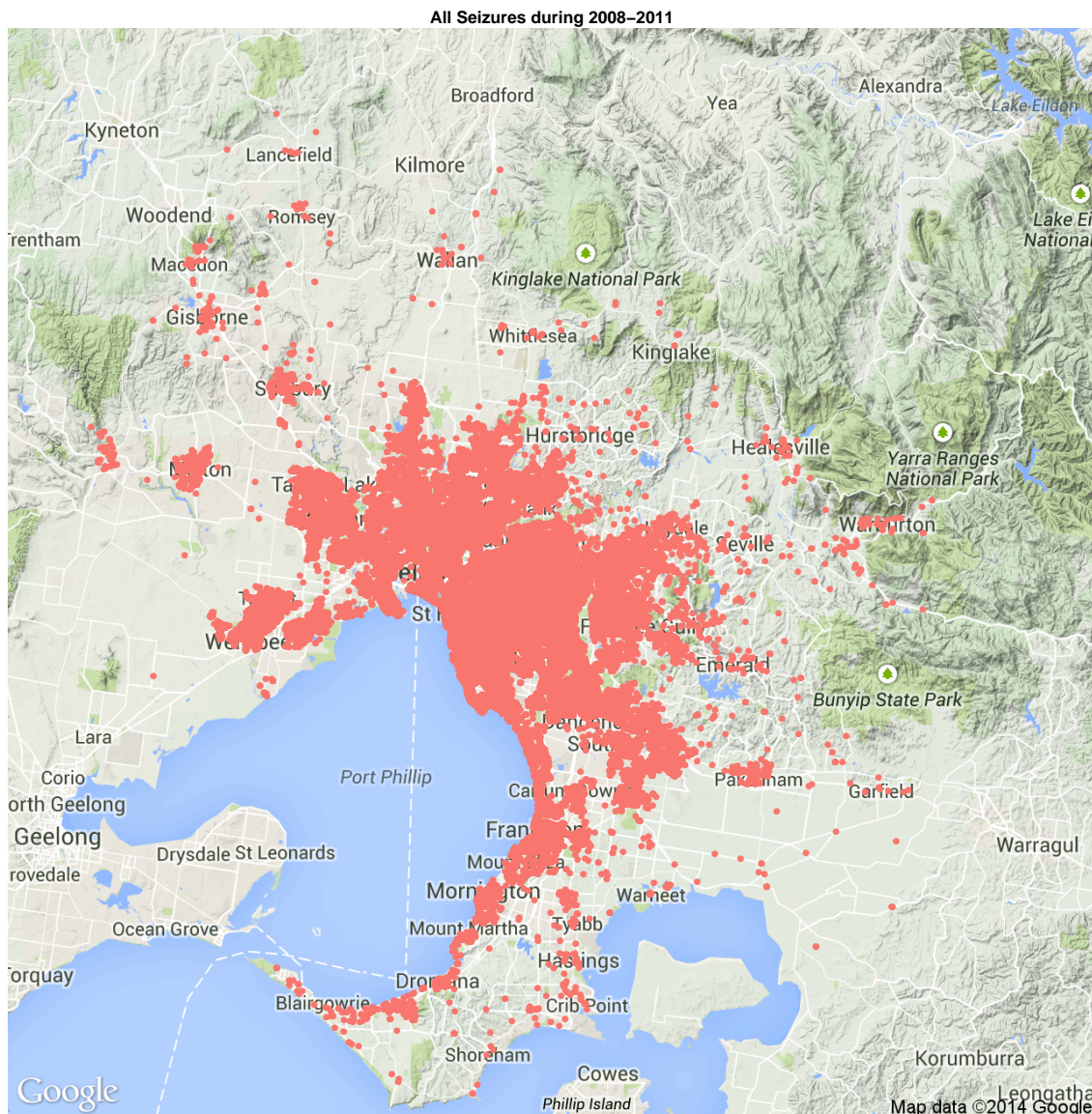
## 2.3 Spatial mapping

This sub-project has involved the provision of geocoding software to enable the department to produce visual displays of various kinds. Hence most spatial mapping has been occurring within the department, and does not feature prominently in this report. However, it may be desirable to supplement this with some spatial maps that can be produced in R, should they be of use in the future.

These maps are provided as illustrative, with the scripts supplied in order for these to be tailored to individual needs.

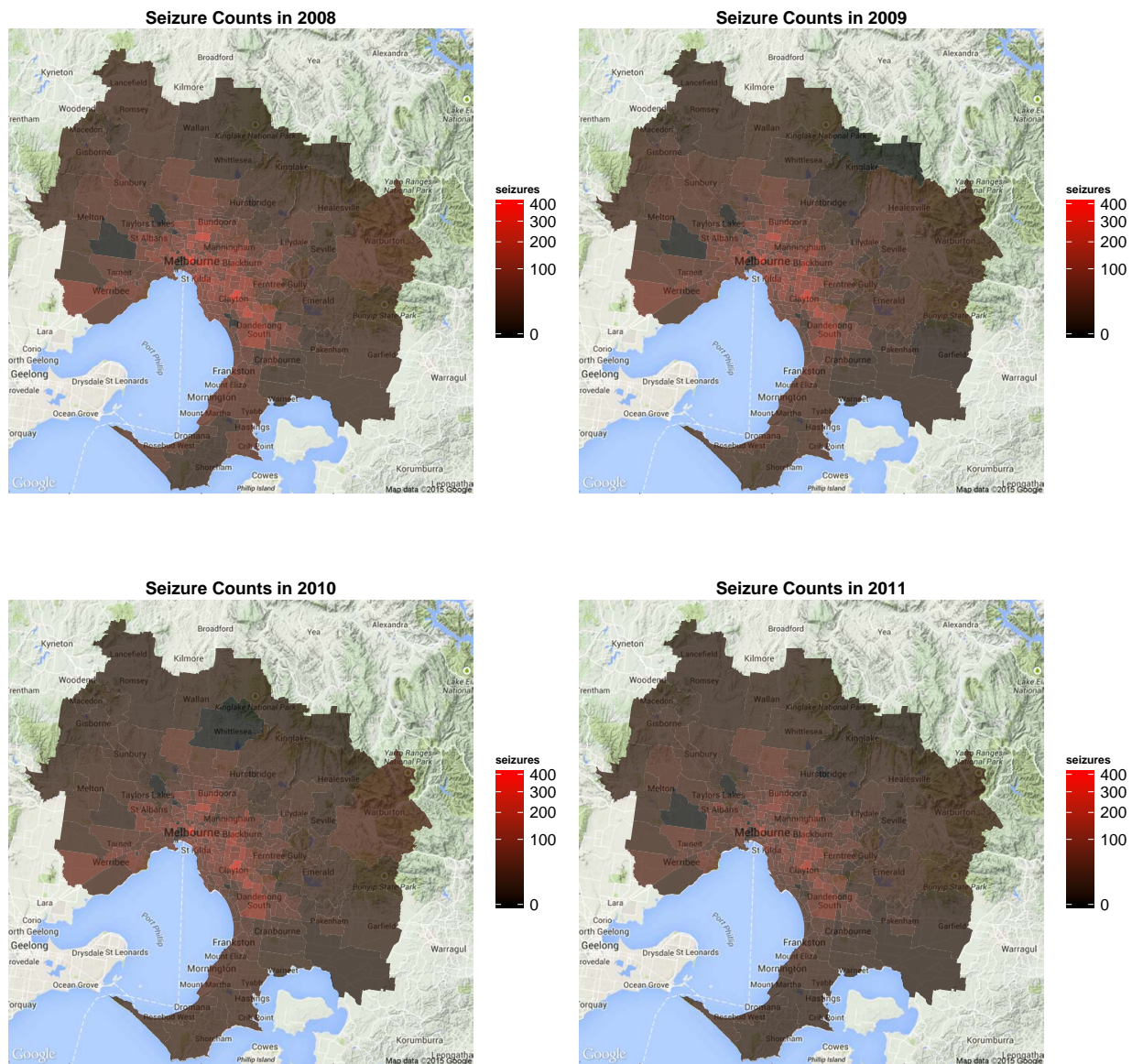
Figure 2.1 shows all seizures in the Greater Melbourne area. Figure 2.2 shows seizures presented by year to demonstrate the change over time, for 2008–2011.

In addition to these spatial plots, the seizures for some key commodities identified by the department team as of particular interest, for Greater Melbourne and Greater Sydney, have been provided to members of this team.



**Figure 2.1:** Greater Melbourne destinations to which seized international mail articles were addressed (raw points) with map.





**Figure 2.2:** Greater Melbourne destinations to which seized international mail articles were addressed (counts) with map.

## 2.4 Relating seizure counts and population characteristics

### 2.4.1 Analysis approach

The primary aim of this sub-project is to relate the demographic data to seizure counts, to determine which demographic fields relate most strongly to high counts of seizures. As the demographic data at the SA2 level were from the 2011 census, we used the seizure information from 1 January 2008 to 26 June 2012. Of the 428,983 seizure records in this period, 326,291 could be matched to a unique SA2 area. Common reasons for a failure to match are that the international sender address was incorrectly included as the recipient address, the address was poorly spelled or the address was fake. Of the 326,291 matched records, 170,914 were found to be misdeclared or not declared and therefore were used for analysis. Declared mail items were excluded: they are not considered a biosecurity risk because, even if biosecurity risk material is present, the declaration should explicitly indicate this. In practice, some declared mail items may in fact be misdeclared and contain unforeseen biosecurity risk material, but it is not possible

to know the extent of this. Of course, declaration cannot be used for screening, but it can be used in this context where the aim is to determine characteristics of high risk area, not screen individual mail items.

The seizure counts can be summarised overall, for each of 12 broad categories, given in Table 2.4, and for those individual commodities with at least 1000 seizures in the time period, given in Table 2.5.

**Table 2.4:** Counts of seizure classes in the mail pathway 2008–2011, sorted by decreasing frequency. This excludes those classified as “Inspected and Released”.

Category	Count
Plant and Plant Products	57747
Animal Products	43195
Human Therapeutics	13099
Grains/Legumes/Nuts	10963
Fruit and Fruit Products	9320
Herbs/Spices	8529
Contaminated Goods/Footwear/Packaging	5954
Vegetable and Vegetable Products	5272
Mushroom/Fungi	4563
Biologicals	1027
Soil/Mineral Samples/Fertiliser	991
Live Animal	119

**Table 2.5:** Counts of seizure types in the mail pathway by key commodity 2008–2011, sorted by descending frequency.

Commodity	Count	Commodity	Count
Dried plant material	19663	Pork	1688
Other plant products	13538	Other therapeutics	1686
Tea	5951	Conifer foliage	1663
Powdered milk/dairy	5116	Khat	1638
Used boxes/cartons	4257	Beans	1622
Other herbs/spices	4114	Noodles/Pasta with Egg	1570
Other seeds	4003	Beef	1530
Salami/sausage/small goods	3904	Other foliage	1530
Other vegetable and vegetable products	3887	Bark	1527
Finfish (non-salmonidae)	3400	Mayonnaise/Egg based sauces	1428
Other mushroom/fungi	3178	Straw	1383
Other live plants	3093	Other fruit/vegetable seed	1350
Other egg products	3027	Walnut	1311
Seasoning with egg	2786	Traditional Medicines Non-Animal	1307
Meat jerky/Biltong	2635	Cloves	1294
Other Fruits and Fruit Products	2258	Pet food/Stockfeed	1277
Unidentified seed	2235	Citrus peel/pomander/leaf	1139
Poultry	2101	Other animal products	1139
Mixed herbs/spices	1898	Almond	1111
Other dairy products	1838	Popping corn	1105
Other meat products	1811		

A flexible model is needed because the nature of the relationship between demographic fields and seizure counts is not necessarily linear nor consistent between different seizure types or demographic data. Tree-based models are valuable in exploratory cases such as this, because they allow for complex relationships between predictors and outcomes, including interactions between predictors in the way they affect the outcome. For the total seizure counts, and each of the separate counts of each category and commodity, we can create a tree which progressively splits each group of observations into two subgroups, based on the level of individual demographics fields. The choice of splits can be optimised automatically



in software to best separate those with high and low potential biosecurity consequences of seizures. The choices of splits yield useful information about the importance of the demographic fields in determining those areas with high seizures counts.

Note that the data in their current form also allow for another kind of interaction by including counts of combinations of languages, education levels, etc. for each SA2, as given in Table 2.3. This data structure permits fitting of cross-tabulation style interaction terms, and enables analysis of the impact of combinations of factors within each SA2, as well as assessing the correlations and interactions between the counts for these combinations at the SA2 level, as modelled directly in a tree model.

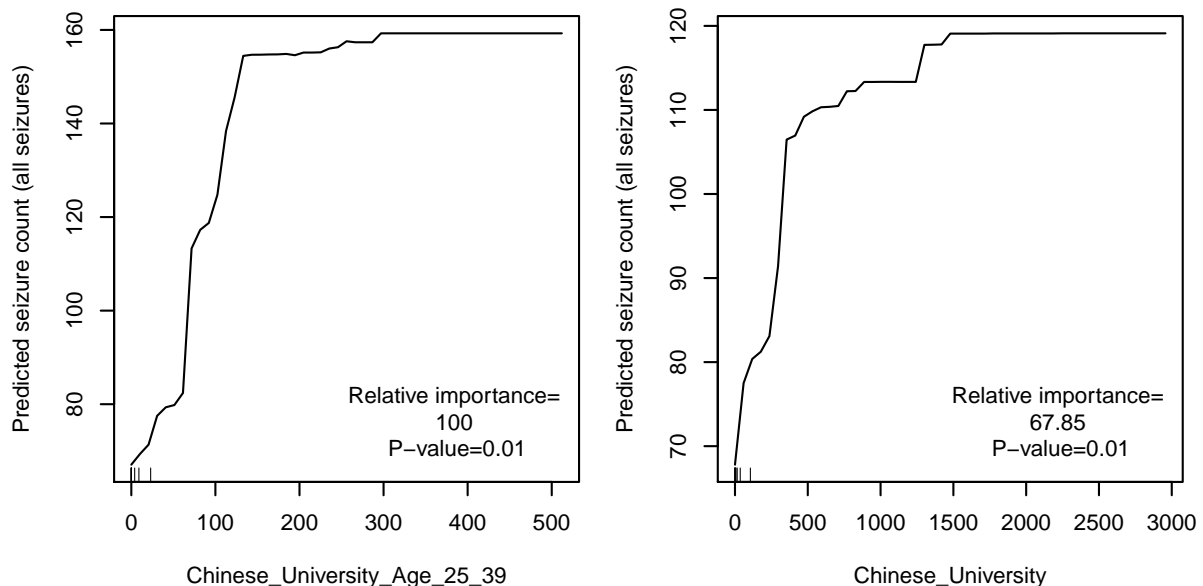
It is possible to construct one such tree model for these data but over-fitting is a genuine risk. As there is a relatively large number of potential demographic predictors available when fitting the model, it is possible to estimate the risk for the existing data very well, but the model may not be generalisable to other pathways. For example, it may highlight spurious relationships due to only a few unusual observations in this particular data set. One way to mitigate this problem is to create many trees based on random samples of the data and average the results of these, so the results are not sensitive to the specific data available. These kinds of models are called random forests, as they involve many trees generated from random samples of the data. The number of SA2s make this approach feasible in this case. More details of this approach can be found in Appendix A.

A different random forest model was constructed for the count of seizures in total, in each of the broad categories and for individual commodities with at least 1000 seizures in the time period. As there were very few SA2 regions for which the seizures for the category of live animal exceeded 1 or 2, these results should not be considered as reliable as for the other categories, however, they have been included for completeness.

The actual implementation involved the `randomForest` package in R (R Core Team, 2013).

## 2.4.2 Description and interpretation of results

As this approach involves the averaging of many tree models, no single tree or diagram can be used to display the relationship between the demographic data and seizure events. However, it is possible to consider the effect of each key demographic field, averaged over the other fields, using overall measures of importance and, visually, using partial plots. These summaries have been provided to the key contacts for this sub-project within the department, for the top 6 demographic fields for each outcome (in total, by category or by commodity). For the purposes of explanation, we will consider the top two fields for the total seizures as an example, given in Figure 2.3.



**Figure 2.3:** Some example partial plots, presenting the estimated overall relationship between two key demographic variables and predicted seizure count, based on the random forest model. The short vertical lines across the base are the deciles of the distribution of the demographic field.

The horizontal axis represents the SA2 count of each demographic field. The range of values on the

axis reflects the range of counts for that demographic field across the different SA2 units. For example, the first of these figures is the number of people who speak Chinese, attend university and are aged between 25 and 39. The second of these is the number of people that speak Chinese and attend university. Obviously these counts are related, but the model has found that both fields are important in the modelling of seizures, and provide important, distinct information.

The vertical axis is the predicted total seizure count per SA2 for the seizure period considered, based on the combined tree models. The line relates the demographic field with the seizure count, based on the various cut-offs generated in the individual tree models. Looking at the first figure, the count of seizures slowly increases initially with increasing numbers of this composite demographic field, increasing sharply at around 80. Once this field reaches around 150 it stabilises. Very small numbers within this category indicate a low seizure risk. In the second figure, the seizure risk also increases with increasing Chinese speaking university students overall, stabilising when this field reaches around 500. It is important to note the range of the vertical axis varies between partial plots; the interpretation of the impact of the demographic data ought to be considered in light of the impact on the absolute count of seizures. Similarly, as the number of Chinese speaking university students aged 25–39 is a subset of the total number of Chinese speaking university students, the range of values that this demographic takes is smaller, and a smaller change in numbers has a greater impact. Also, while these two examples indicate a generally monotonically positive relationship, there is nothing in the modelling process that requires such a relationship and some negative or non-monotonic relationships have been observed. It is just that the kinds of predictors chosen for consideration in this modelling have been those expected to be positive related to the count of seizures.

Each figure reports the relative importance as given by the random forest model, and a  $p$ -value for the statistical significance of that importance value. These are a summary of impact of the splits based on each field. These importance values are only relative and have been standardised such that the largest value is set to 100. This gives an indication of the relative importance of each demographic field in the prediction of seizures, as compared with the relative importance values for the other demographic fields reported for the same model. The  $p$ -values are based on a comparison with models with the outcome randomly permuted. These are obtained using the `rfPermute` package in R (R Core Team, 2013). For computational simplicity, only 50–100 such random models were used, resulting in a discrete range of possible  $p$ -values and therefore repetition of these across the demographic fields.

Note also the short vertical lines at the bottom of the graphic. These give an indication of the distribution of the demographic fields, as these are deciles of the distribution which divide the SA2s into 10 equally sized groups. As these indicate, all these demographic fields are highly skewed, with the vast majority of SA2s having very small numbers in the demographic categories. Some other demographic fields are not as concentrated to as few SA2s, such as the overall age or language categories, rather than the composite fields which tend to be more sparse. Some of the vertical lines are so close that they form a bar, and some of the groups are identically zero.

It is important to clarify at this point that all seizure outcomes were based on the absolute counts of seizures. As approach counts for each SA2 were not available, it is not possible to accurately calculate seizure rates. Total population counts for each SA2 were included in the model, but were not found to be importance predictors. This suggests that the important predictors for an analysis of the rate per person at each SA2 would be similar.

Scripts for implementation in R have been provided which generate figures like these for the top 12 demographic fields, as well as the full set of relative importance values for the demographics fields considered in the model.

### 2.4.3 Summary of key results

The demographic fields most commonly related to high overall seizure counts involve Chinese language, persons aged 15–39, and education either: university, not applicable (that is, not attending an educational institution) or not stated. These demographic fields reoccur for specific categories of seizures and seizures of key commodities, with some additional demographic fields relevant in individual cases.

Focusing on the broad categories of seizures in Table 2.4, high levels of seizures of animal products seemed to occur in areas with high levels of some of the other Asian languages. High levels of Baltic languages were related to seizures of biologicals, and high levels of European languages related to soil/mineral samples and fertilisers. In places with high levels of seizures of human therapeutics, there was an observed relationship with increased amounts of African languages, and seizures of live animals were related to South East Asian languages, as well as some European, African and other Asian languages.

This relationship was reported in the model and therefore considered statistically important.

Focusing on the key commodities in Table 2.5, again, most seizures of these were related to Chinese language, persons aged 15–39 and education either university, not applicable or not stated.

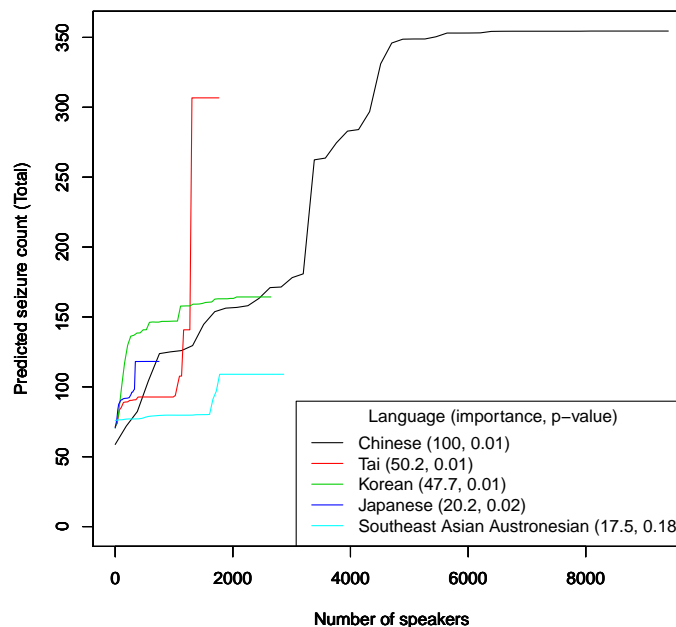
However, conifer foliage was particularly related to Scandinavian and French speakers, along with a range of other European and Asian languages. Similar language associations were observed for seizures of cut flowers, other foliage, other live plants and popping corn. Seizures of tea and straw seem to occur more in regions with higher levels of French and Japanese.

For finfish, higher levels of Korean and Tai (NB: *not* Thai) languages were strongly related, and the prevalence of African languages was unsurprisingly by far the strongest predictor of high counts of seizures of khat. High numbers of Japanese speakers were related to mayonnaise/egg based sauces, noodles/Pasta with egg and seasoning with egg, with these also related to Tai languages. Other fruit/vegetable seeds were related to Middle Eastern Semitic languages, as well as the number of Chinese speakers working in Mushroom/Veg and Poultry related industries.

Salami/sausage/small goods were harder to analyse, with the number of people with language *not stated* a good predictor, along with Tai, Japanese, Scandinavian and French languages. That is, there was no evidence of clean ‘continental’ distinctions. Seizures of other seeds and pet food/Stockfeed-animal derived seem to relate only to overall population size (increasing seizures with increasing SA2 size) aside from some Japanese and East Slavic language associations with the latter.

#### 2.4.4 Additional results

In addition to the results of models with all demographic fields considered, there was also an expressed desire to focus on the languages only to understand the relative strength of the predictive power of each language group. Figure 2.4 provides the results for the top five languages in terms of importance in predicting total seizures, for a random forest model constructed using the 60 languages. Each language has a relative importance as given by the random forest model, and a  $p$ -value for the statistical significance of that importance value. The language curves stop at different points because they correspond to different numbers of seizures ( $y$ -axis) and numbers of speakers ( $x$ -axis).



**Figure 2.4:** A partial plot for the top five languages for predicting total seizures within a census region in the mail pathway.

## 2.5 Conclusion and Recommendations

The results of the model are complex to describe, but the model can be used to make predictions for a given situation. That is, for a given demographic breakdown, this model can provide an estimate of the count of seizures. As this particular application pertains more to the development of targeted campaigns, determining the key demographic fields as discussed above is going to be more useful than being able to predict the count of seizures for a given hypothetical or future situation. The random forests data analysis shows that both attendance at University and speaking languages other than English are strong predictors of greater likelihood of receiving mail articles with BRM. It is reasonable to conjecture that food parcels are being sent to foreign university students by family or friends, who may be ignorant of Australia's strict biosecurity regulation.

**Recommendation 1.** Targeted public relations campaigns may reduce the incidence of biosecurity risk material being sent in the mail. Foreign-language university student associations should be asked to disseminate relevant biosecurity information to their members, and to assist in identifying legal sources of regulated goods that are commonly intercepted in the mail or air cargo pathways. A before-after control-intervention design should be used to assess evidence for the conjecture.

The exercise of applying spatial statistical tools to interception data for the mail pathway has led to several further directions, including specific potential case studies for further analysis, such as seizures of tea, seeds, khat, and finfish. It is likely that the techniques used in this chapter will also be useful for profiling in the air cargo pathway. Air cargo documentation includes delivery addresses, and has the potential benefit of allowing profiling upon arrival because address information is recorded electronically.

**Recommendation 2.** Regulated goods are transported by air and sea cargo as well as by international mail. Spatial analysis similar to the mail analysis should be applied to destination addresses of seized articles in the air and sea cargo pathways. Cargo addresses are captured electronically, so the analysis will be more straightforward, and profiling of geographical hotspots for intervention would be possible.

Predictive modelling will become more important should the department wish to develop their screening approach to target individual mail items. This would require consideration of other characteristics of the mail items, such as country of origin and the description, and any modelling would require the inclusion of information about the mail items that were not seized, as negative controls. As identified by Claire McKee, Australia Post has implemented a system in the Sydney Gateway Facility that may enable counts of the total count of mail items delivered to each postcode. This information would then permit profiling at the postcode level by estimated approach rate.

**Recommendation 3.** The Sydney Gateway Facility has equipment that counts the international mail articles delivered by postcode. For Sydney data, obtain post-code specific delivery rates to estimate the relative risk of deliveries to specific locations and determine whether any are sufficiently risky to justify post-code targeting, as is done by the Australian Department of Immigration and Border Protection.

Another potential future direction for this sub-project identified by Claire McKee is the use of the department's land-use classification for each postcode. These are available via the '2005/2006 Land Use of Australia' data set, version 4. If available, this could supplement any risk modelling, enhancing the interpretations of results.

**Recommendation 4.** The material risk of contamination transported by international mail may depend on the nature of its destination. Use the post-code of the delivery address to distinguish between urban and rural locations, and assess any effect upon the risk of contamination, from the points of view of (i) approach rate and (ii) magnitude of consequences.

Unrelated to the aims of this sub-project, which has focused on known misdeclared mail items, reviewers suggested that some declarations within MAPS are unreliable. The extent of this problem is unknown, so a project designed to explore and improve on the use of this particular field may be valuable.

**Recommendation 5.** The declaration field on the international Customs form for international mail articles is hand-coded by the sender or their representative. Various factors cast doubt on the value of this information. Assess the veracity and utility of the international mail declaration field on the Customs form.

## Chapter 3

# Generalised Pattern Analysis for International Passengers

SANDY CLARKE<sup>\*</sup>, ALAN KÜFFER<sup>†</sup>, AND ANDREW ROBINSON<sup>‡</sup>

### 3.1 Summary

#### 3.1.1 Background

This chapter provides an update to the results in the profiling of passengers as part of the Generalised Pattern Analysis sub-project. The previous analysis found that there was strong predictive value of the Border data for identifying passengers who were found to be carrying biosecurity risk material, but noted that the intervention information was not available. The analysis reported in this chapter incorporates intervention type for the analysis (including release upon documentation, bag x-ray, K9 or manual inspection, as per the recommendations of Clarke et al. (2015a)).

#### 3.1.2 Motivating Question

Briefly, the project seeks to predict the compliance of inspected passengers by using pre-arrival data provided by the Department of Immigration and Border Protection (DIBP). Here, *compliant* means that the passenger was not discovered to be carrying any Biosecurity Risk Material (BRM).

If the compliance of the passengers could be predicted well by DIBP data, then two advantages would be possible. First, the passengers could be profiled before arrival, so could be more efficiently directed to intervention at the primary line. Second, passengers could be profiled at the flight level, which would enable more efficient resourcing.

#### 3.1.3 Methods

The data resources were a selection of passengers from Adelaide and Sydney for which an Incoming Passenger Card (IPC) had been coded, in order to determine the intervention type. The intervention type was extracted by interpreting the standardized quarantine mark-up used at each region. This information was combined with auxiliary passenger data along with compliance information, as in Clarke et al. (2015a). There were 4,685 such records for Adelaide and 64,903 records for Sydney.

We used a statistical technique called a *gradient boosting machine*, which takes many random samples of observations and predictors and fits the best models it can to these samples, and then using them to build an averaged model. The samples are weighted so that observations that are found to fit poorly are subsequently more highly weighted. Before modeling, we thinned the candidate predictors. These steps are documented in more detail herein.

---

<sup>\*</sup>Statistical Consulting Centre, The University of Melbourne

<sup>†</sup>Compliance Division, Department of Agriculture and Water Resources

<sup>‡</sup>CEBRA, The University of Melbourne

### 3.1.4 Results

Overall the results are promising. The DIBP data showed a strong relationship with the outcome of the passenger inspection, but we must note that the sample of passengers inspected was not random because the passengers not flagged by the original biosecurity protocol were not inspected. Therefore the results should be interpreted with caution.

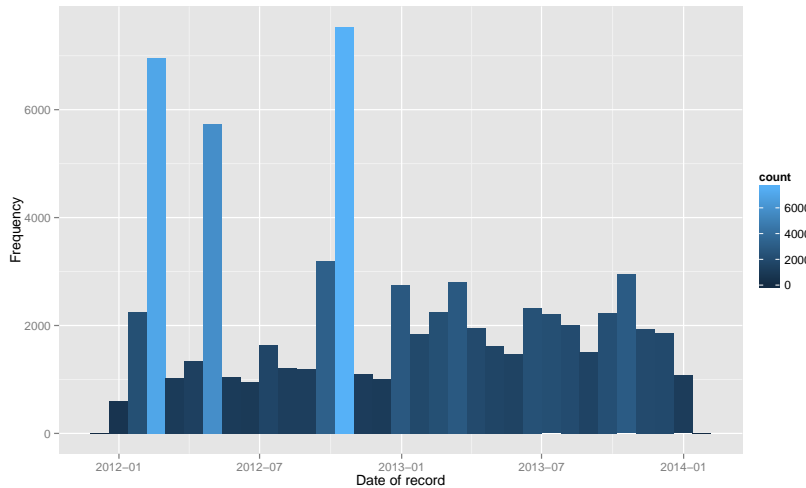
### 3.1.5 Conclusions and Future Directions

In short, the early signs are promising, but considerable work needs to be done with a more representative sample before firm conclusions can be drawn. The following short-comings of the present study need to be remedied, if the potential benefit seems worthwhile.

## 3.2 Data preparation

The data resources were a selection of passengers from Adelaide and Sydney for which an Incoming Passenger Card (IPC) had been coded, in order to determine the intervention type. The intervention type was extracted by interpreting the standardized quarantine mark-up used at each region. This information was combined with auxiliary passenger data along with compliance information, as in Clarke et al. (2015a). There were 4,685 such records for Adelaide and 64,903 records for Sydney.

These records were supposed to be random samples of the IPCs in 2012 and 2013, however there are some indications that this is not the case. Characteristics of this sample differ greatly from the full set of passenger records from the period. For example, Figure 3.1 indicates a very uneven distribution of records across the time period, with cards from passengers arriving on some individual dates (2012-05-02, 2012-10-23 and 2012-02-15) occurring 10 to 100 times more frequently than other dates. There are marked differences in the travel document countries, as indicated in Table 3.1, with some more minor differences in the flight numbers, as indicated in Table 3.2. The impact of these differences on the modelling results is difficult to assess. Passengers with travel documents from Australia and New Zealand were excluded from the analysis.



**Figure 3.1:** Distribution of international passenger inspection records over time.

The full set of fields considered for the purposes of modelling are given in Table 3.3, with a brief description and summary of data cleaning process used, as in Clarke et al. (2015a), but now with the addition of intervention type to allow for the different interception rates that arise from different interventions.

The next step was to reduce some of the categorical fields with many levels, including some with very small counts. Where a level of a categorical field featured less than 20 times, these categories were collapsed into one category called ‘small’. The fields for which this was necessary were:

- Check\_in\_Port\_Code;
- Travel\_Doc\_Dept\_Country\_Code;

**Table 3.1:** Comparing the frequencies of the top 10 travel document countries with their frequencies in the IPC sample.

Country	Total passengers	IPC sample
AUSTRALIA	48.5%	0.0%
NEW ZEALAND	8.8%	0.0%
CHINA	7.0%	14.0%
UNITED KINGDOM	5.8%	15.3%
UNITED STATES	4.4%	11.6%
KOREA	2.4%	5.3%
JAPAN	2.2%	5.0%
INDIA	1.9%	2.6%
GERMANY	1.6%	3.6%
MALAYSIA	1.5%	3.0%

**Table 3.2:** Comparing the frequencies of the top 10 flight numbers with their frequencies in the IPC sample.

Flight number	Total passengers	IPC sample
EK412	2.1%	1.8%
EK414	2.1%	2.1%
EK418	2.0%	1.5%
EY454	1.9%	1.4%
EY450	1.8%	0.9%
EK413	1.8%	0.9%
EK419	1.7%	0.8%
SQ279	1.3%	1.5%
QF82	1.3%	0.9%
BA9	1.3%	0.3%

- `Birth_Dept_Country_Code`;
- `Route_ID`; and
- `Visa_Sub_Class_Code`.

Note that this collapsing also enables future prediction on any new, rare levels to be modelled, by treating them as if they were in this `small` category. There is no guarantee that such predictions will necessarily be useful.

Some fields required specific collapsing. The full `Local_Scheduled_Date` was not used in the model, but instead was used to create two fields: `Month` and `Year`. In addition to the collapsed `Route_ID` described above, a separate field with the first two letters of `Route_ID`, which correspond to the airline, was also created and underwent the standard collapsing of levels with counts less than 20.

In general, all missing or blank cells were replaced with a missing value flag (`NA`) once it was confirmed there were no clashes between this and any existing codes used. This could then be considered a level in its own right.

For the purposes of analysis, all categorical fields were converted into a series of dummy fields, one for each level (1 indicating that level was present for the given record and 0 otherwise), to reduce the data complexity and separate the effects of individual levels more clearly.

### 3.3 Data analysis

The outcome for modeling is the event that the passenger is detected as carrying biosecurity risk material. An initial reduction of fields (also known as feature selection) was performed, before a formal model with the remaining fields and their interactions could be considered. In order to do this reduction, the dataset was randomly divided into two halves, a training set and a test set. The training set was used to fit a simple generalised linear model which was then used to predict the outcome for the test set. The

**Table 3.3:** Data fields available for international passenger risk analysis.

Field name	Description	Notes
Seizure	Indicator variable of non-compliance	No data cleaning required
Sex_Code	One character code for gender	Sparse levels collapsed
age	Traveller age on arrival	No data cleaning required
Month	Arrival month	Derived
Year	Arrival year	Derived
Local_Port_Code	Three character code for Australian arrival airport	No data cleaning required
Route_ID	Parent flight number	Sparse levels collapsed
Airline	Airline Code	Derived from <code>Route_ID</code>
Check_in_Port_Code	Three letter check in airport code	Sparse levels collapsed
Passenger_Crew_Code	One letter crew or passenger indicator	Sparse levels collapsed
Visa_Sub_Class_Code	Visa type for foreign passport holders	Sparse levels collapsed
Travel_Doc_Dept_Country_Code	ISO alpha 3 standard country of citizenship code	Sparse levels collapsed
Birth_Dept_Country_Code	ISO alpha 3 standard country of birth country code	Sparse levels collapsed
total_trav_count	Cumulative count of trips at point in time	
period_of_stay	Time outside of Australia	
Endpoint	Inspection type	Collapsed into four types of screening

performance of this was assessed using the area under the Receiver Operating Characteristic (ROC) curve, AUC, for the test set; a common means for assessing predictive performance.

An ROC curve is a plot of the proportion of the non-compliant passengers that have been detected (y-axis) against the proportion of the inspected passengers that are compliant (Hastie et al., 2009); for a biosecurity-related application see Robinson et al. (2016). We can use an ROC curve to assess models by undertaking the following steps for each model: sort the cohorts into descending order of contamination rate and plot the cumulative number of non-compliant articles against the cumulative number of detected compliant articles. AUC is then the size of the area between the ROC and a line at  $y = 0$ , evaluated from  $x = 0$  to  $x = 1$ . AUC range from 0 to 1, and an AUC of 0.5 implies that the profile from which the curve is derived is no better than a random guess.

This process was repeated 5 times and summarised by the average AUC. Any field which had an average AUC greater than 0.510 was carried over into the next analysis step; note that an AUC of 0.5 or less indicates no predictive power. Fields were considered independently of each other at this stage.

Table 3.4 contains the list of those fields and levels that were carried into the next phase of analysis, 57 in total, because of evidence of individual predictive power. Note that predictive power doesn't only imply greater risk; this also includes fields that show notably lower risk of the outcome.



**Table 3.4:** Data fields for international passenger risk analysis after initial feature selection.

Field name	Levels
Sex_Code	female male
age	numerical field
Month	January
Year	2011 2012
Local_Port_Code	ADL CNS OOL PER SYD
Route_ID	CX105 CZ321 CZ325
Airline	CA CX CZ DJ EK MH MU NZ QF SQ
Check_in_Port_Code	small AKL CAN CHC DPS HKG LAX LHR PEK PVG WLG
Passenger_Crew_Code	crew passenger
Visa_Sub_Class_Code	small 456 573 676 942 NA
Travel_Doc_Dept_Country_Code	AUS CHN GBR IND NZL
Birth_Dept_Country_Code	AUS CHN GBR IATX IND NZL VNM
total_trav_count	numerical field
period_of_stay	numerical field

Now that the set of predictor fields had been sufficiently reduced, a model combining these fields could be used. The same approach was applied as is documented in Clarke et al. (2015a), and is included here for completeness.

Now that the set of predictor fields had been sufficiently reduced, a model combining these fields could be used. The model approach used involved decision trees with boosting and shrinkage, using the gradient boosting machine package (**gbm**) in R (R Core Team, 2013). More details of this approach can be found in Appendix A. Tree based models progressively split observations into two groups, where the splits are based on levels of the predictors. Tree-based models are valuable in exploratory cases such as this, because they allow for complex relationships between predictors and outcomes, including interactions between predictors in the way they affect the outcome. This is required because the nature of the relationship between predictors and outcomes is not necessarily linear nor consistent between different predictors.

The boosting and shrinkage aspect involves fitting these tree models many times, for each new step, weighting those observations that were hardest to classify correctly in the previous step. In addition to this automated re-weighting, this model includes an initial proportional weighting of the number of non-compliants, to reflect the fact that there were ten times as many compliants in the sample used.

This is a reasonably computationally intense method, but can still be implemented on a standard desktop if the initial data reductions discussed above are applied. Should further computational reduction be required, the **randomForest** package within R might be a useful alternative. This involves the averaging of many trees, generated from random samples of the data, and provides results in a very similar format.

The methods used here draw on Miller et al. (2009).

### 3.4 Results

As this approach involves many tree models, there is no single tree or diagram that can be used to display the results. However, it is possible to consider the contribution of each predictor, averaged over the other fields, using overall measures of importance. Table 3.5 provides a list of these relative measures of importance, ordered by importance, for the first 16 field/level combinations. These importance values are. They are only relative and have been standardised such that the largest value is set to 100.

Each row also contains an odds ratio for the relative odds of non-compliance for passengers with that variable level compared to those without it. For example, if the birth country is not Australia, the overall odds ratio is estimated to be 8.33; that is, the odds of non-compliance if the passenger is not from Australia are more than 8 times higher than passengers born in Australia. Even though we have increased the relative number of non-compliants in the sample used to generate the model, this is a consistent estimate of the odds ratio for the population. Note that some levels, like this one, are expressed negatively, if the effect is to reduce the risk.

The final column of Table 3.5 gives the rarity of this field level, which is the percentage of data records with this variable level. Where the variable level is expressed in the negative, the rarity is the percentage of records without that level.

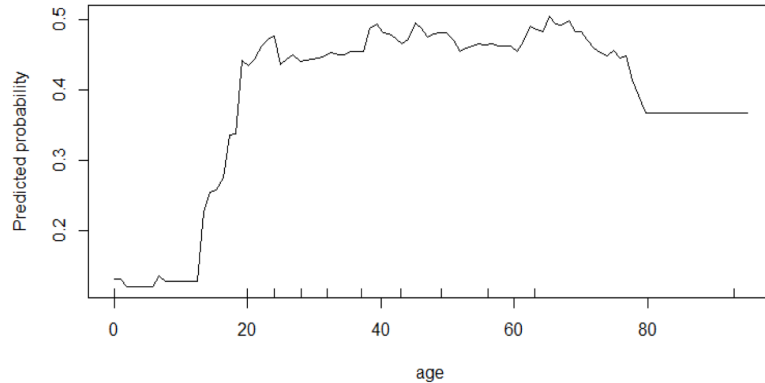
This does not provide an indication of the nature of any interactions between fields in the model, though undoubtedly such effects are going to be involved. Such effects are included in the model itself and any prediction arising from the model.

Each row also contains an estimated odds ratio for the relative odds of non-compliance for passengers with that variable level compared to those without it.

**Table 3.5:** Relative importance and crude risk for international passenger data, for the top 10 field levels.

Field and level	Importance	Odds Ratio	Rarity
Screening not None	100	22.4	35.8%
age	12	NA	NA
Visa_Sub_Class XXX	5	2.79	4.2%
Screening Xray	5	6.01	10.7%
Birth_Dept_Country_Code XXX	5	3.04	14.8%
Local_Port_Code XXX	4	0.44	6.7%
Screening Manual	4	1.51	12.6%
Local_Port_Code XXX	3	2.26	93.3%
Travel_Doc_Dept_Country not XXX	3	5.04	84.7%
Travel_Doc_Dept_Country XXX	2	1.56	4.2%

Although it is possible to report relative importance for numerical fields, it is not possible to give one overall risk measure because the fields can take a range of numerical values. Instead, we can assess the effects of these visually, using partial plots, for example as provided for age in Figure 3.2. As this figure indicates, there is increased risk for adults than children, although this effect is not as strong as was observed in Clarke et al. (2015a). The attenuation of the signal compared with the previous study is likely due to the inclusion of intervention information in the currently reported analysis.



**Figure 3.2:** The relationship between age and the predicted probability of non-compliance for international passengers. A rug plot is added to provide information on the distribution of age.

The clear message from this summary is the strong association between inspection type and the detection of non-compliance, particularly whether or not screening was performed. This result is confirmatory given the greatly different rates at which the differently screened passengers are searched. The relative importance of all other field levels are much smaller in comparison.

To understand the impact in terms of overall predictive performance, we partitioned the data into five equal sized subsamples. For each subsample of 20% (test data), we repeated the modelling approach including feature selection and gradient boosting trees using the remaining 80% of the data (training data), in order to leave 20% the test data for assessing the model performance.

If we take those predictions above the 75% quantile (that is, the top 25% in terms of risk as predicted by our model), and compare these to their actual compliance, we get the results in Table 3.6. For this approach, we observe a sensitivity of 76.8% and a specificity of 75.5%. The odds ratio for this model gives an overall measure of its performance; the odds ratio in this case is 10.2 (95% CI: 8.4,12.4). The performance of this model is slightly worse than that documented in Clarke et al. (2015a). Given that this model used the same predictors plus the addition of another, very strong predictor, this performance is surprising. There are a few possible reasons for this which include the seemingly selective nature of the sample, and the overall small number of seizures available for the model training.

**Table 3.6:** Performance of the full model on withheld passenger inspection data (with proportions).

	Compliant	Non-compliant
Predicted compliant	52049 (0.75)	142 (0.002)
Predicted non-compliant	16927 (0.24)	470 (0.007)

It is important to note that the results in this Table are drawn from models that include intervention type as a predictor. Passengers who are released upon the provision of suitable documentation are very unlikely to be discovered carrying biosecurity risk material.

### 3.5 Implementation Guidelines

Scripts for fitting the models discussed so far have been provided to Agriculture. This is to ensure that models can be adapted both to changes in available data (such as new records or changes in flight codes) as well as changes in the implementation needs. For example, the current model suggests that the year of the record is of some importance, which may be an interesting finding, but is not likely to be useful for future profiling. The implementation of any approaches proposed ought to be iterative, with regular reassessment of performance.

One likely use of these results is the provision of predictions for a given set of passenger characteristics. The results above indicate which fields are important in this prediction, but do not provide a model for prediction as such. The `gbm` package can provide predictions, in the form of a estimated probability of non-compliance. As the model used is complex, and includes interactions, it cannot be reduced to a simple closed form equation, but predictions for a set of new data can be made using R. However, real-time profiling may need to occur independently of R depending on the hardware used in implementation this.

One solution is to provide a “look-up” table with predictions for all the combinations of predictors. However, the number of combination may be too large for this to be feasible. Even without the numerical fields, which take a very wide range of values, there are trillions of combinations of the binary fields which could be accounted for.

We could choose to focus on a few of the key fields, and fit a model to these only. For example, prediction results for a model involving only these four field levels:

- `Birth_Department_Country` AAA or BBB;
- `Travel_Doc_Dept_Country` CCC; and
- `Visa_Subclass` NA.

Although the full data were used to assess the risk factors and their relative importance, in order to best assess the performance of the model, it is appropriate to separate the data into testing and training data. We fit the simple model above based on a random 80% of the data (training data), in order to leave 20% of the data (test data) to use for assessing the model performance.

Results for this model based on 80% of the data are given in Table 3.7, in the kind of format that could serve as a look-up table.

**Table 3.7:** Predictions for the simple passenger profile model using four different field levels.

Field and level						
Birth_Dept_Country_Code AAA	0	1	0	0	1	0
Birth_Dept_Country_Code BBB	0	0	1	0	0	1
Travel_Doc_Dept_Country_Code CCC	0	0	0	1	1	1
Visa_Sub_Class_Code NA	0	0	0	0	0	0
Predicted probability	0.559	0.038	0.828	0.25	0.047	0.632
Field and level						
Birth_Dept_Country_Code AAA	0	1	0	0	1	0
Birth_Dept_Country_Code BBB	0	0	1	0	0	1
Travel_Doc_Dept_Country_Code CCC	0	0	0	1	1	1
Visa_Sub_Class_Code NA	1	1	1	1	1	1
Predicted probability	0.512	0.135	0.694	0.119	0.105	0.328

The nature of the bias for the input sample means that these predictions will be inflated in probability, but the relative differences are still informative.

We can test the performance on this model for the 20% test data that were not used to generate the model. For example, if we take those predictions above the 75% quantile (that is, the top 25% in terms of risk as predicted by this simple model), and compare these to their actual compliance, we get the results in Table 3.8. This is equivalent to sampling 25% of passengers, based on estimated risk. This corresponds to a sensitivity of 35.7% and a specificity of 92.5%, and an odds ratio of 6.8 (95% CI: 5.9, 8.0). It is clear the model itself is overly simplistic and does not perform as well as more complex models, but this simple example serves to illustrate the potential with limited real-time profiling algorithms.

**Table 3.8:** Performance of simple predicted passenger profile model.

	compliant	non-compliant
predicted compliant	8796	627
predicted non-compliant	713	348

Note that, in order to incorporate a numerical field into simple models like this, some categorization of the numerical fields may be used. For example, for a field like `period_of_stay`, the prediction plots indicate distinct steps in the risk, so the risk prediction based on this field could be broken up into groups according to these steps. The relevance of this depends on the limitations of implementation, so it is preferable at this point to model with the richest data available.

### 3.6 Conclusion and Recommendations

This analysis has been performed using a non-representative sample of IPCs taken from Sydney and Adelaide international airports. The sample did not include passengers of Australian citizenship, which may have reduced the statistical strength of the signal. Furthermore, the only inspection data available was for passengers that had been selected for inspection by some means, a process that reflects the current screening practices. It is impossible to say whether screening that was based on passenger information supplied by Border would capture a higher or lower proportion of non-compliant passengers. Finally, the model reporting is based on inclusion of predictor variables that reflect operational history rather than available information. Therefore, any conclusions drawn from this analysis must be regarded as preliminary.

In short, the early signs are promising, but considerable work needs to be done with a more representative sample before firm conclusions can be drawn. The following short-comings of the present study need to be remedied, if the potential benefit seems worthwhile.

**Recommendation 6.** The potential benefits and challenges of international passenger screening using Australian Department of Immigration and Border Protection data should be assessed in terms of screening decision timeliness. Specifically, (i) will Border be willing and able to routinely make passenger data available to the department? and, (ii) will there be some means of flagging passengers for quarantine intervention at the primary line?

**Recommendation 7.** Conditional on a positive outcome for Recommendation 6, develop a representative sample of passenger records for which the screened intervention is known and develop analytical techniques to better handle the different interception rates associated with the different intervention types.

**Recommendation 8.** Conditional on a positive outcome for Recommendation 7, develop a snapshot survey using profiles developed from the screening model fitted to Immigration and Border Protection data, and inspect passenger cohorts rated as higher risk by either the Border data profiles or the departmental profiles. Analyse and compare the effectiveness of the profiles.

## Chapter 4

# Detecting Anomalous Broker Activity

SANDY CLARKE\*, NICHOLAS CLARK†, AND DAVID FLEMING†

### 4.1 Summary

This chapter reports analyses of declaration amendment behaviour following quarantine directions, as part of the Data Mining sub-project: Detecting anomalous broker activity.

#### 4.1.1 Background

Brokers are a critical actor in the importation of most goods to Australia. Brokers are intermediaries between the exporters, the regulator, and the importers. Much of the goods that present biosecurity risk will have had their associated transactions, which are on the Integrated Cargo System (ICS) and the Agriculture Import Management System (AIMS), processed by a broker.

The following description assumes that the imported goods are handled by a broker. Details for a FID (Full Import Declaration) are initially provided to ICS by the broker. The information that is registered with the declaration is used to *profile* the goods for biosecurity risk. Examples of the kinds of information that the profiles use are economic tariff number, and the answers to community protection (CP) questions. If the goods information corresponds to one or more of a large number of automated rules, then the FID is lodged in AIMS by ICS, and AIMS issues directions for the submission of further documents such as treatment certification, or directs the goods to quarantine upon arrival for inspection. Among the possible directions, we consider some of them *threatening*, that is, they are likely to signal biosecurity interest, and among the threatening directions, we consider some to be *special*, in that they portend direct biosecurity intervention.

At this point, sometimes the broker issues an amendment to the FID. The amendment might be as simple as correcting the spelling of the delivery address, or it may be a change in the description of the goods. In some cases the alteration may change the biosecurity status of the FID, in which case the FID may no longer be processed by the biosecurity regulator.

#### 4.1.2 Motivating Question

The question that motivates this chapter is: is there a relationship between whether a direction upon a FID is seen to have biosecurity implications — a threatening or special direction — and the probability that the FID is then amended? And if so, is there any relationship between the type of direction and the type of amendment?

These questions are jointly pointing to the higher-level question of whether amendment after a threatening direction could be seen to present an increased threat of biosecurity risk. The approach we are taking is indirect, because the goods for which the FID was amended to avoid biosecurity intervention are,

---

\*Statistical Consulting Centre, The University of Melbourne

†Compliance Division, Department of Agriculture and Water Resources

naturally, not inspected. The outcome of this exercise might reveal whether it would be worth imposing an endpoint survey upon such goods.

### 4.1.3 Methods

We collected from ICS and AIMS all the directions and amendments for 20 brokerages and 13 importers during the period 1 January 2013 and 31 December 2013. There were 79,183 directions in total, associated with 93,535 declarations and 1,283,445 sets of CP questions and answers. We computed odds ratios to determine the magnitudes of the relationships between directions and amendments.

### 4.1.4 Results

1. There was a weak but significant relationship between threatening directions and overall amendments, but there were no specific amendment types that were significantly associated with threatening directions.
2. Threatening directions do not appear to be associated with more substantial amendments; there is, if anything, a tendency towards less serious amendments, particularly for the number of tariff class changes.
3. The overall rate of amendment is higher for special directions than for threatening directions; an increase in overall odds of amendment of 17% for special directions is nearly half again higher than 12% for threatening directions.
4. Curiously, special directions result in significantly *fewer* tariff class changes compared to the remaining directions.
5. Special directions demonstrate very similar patterns to threatening directions: no particular difference or a tendency towards less serious amendments, particularly for the number of tariff class changes.
6. Very few amendments resulted in the introduction of a high-risk CP answer, but a substantial number resulted in the removal of any high-risk answers. *A quarter of all amended FIDS that originally had high-risk CP status were changed to low-risk CP status.*

### 4.1.5 Conclusions and Future Directions

The results show high sensitivity of CP status to the directions. Other amendments do not show the same sensitivity. Further analysis may develop actionable intelligence, and should be considered.

## 4.2 Data preparation

The initial data provided for this sub-project were directions and amendments for 20 brokerages and 13 importers during the period 1 January 2013 and 31 December 2013. This was a targeted sample based on previous campaigns and assessments. Specifically:

- Data from the Integrated Cargo System (ICS) providing all the import declarations lodged by the 33 entities noted above in the time period.
- Data from ICS with all the Community Protection (CP) questions and answers, for the declarations in the scope: 1,283,445 sets of questions and answers.
- Data from the Agriculture Import Management System (AIMS) providing all the quarantine directions applied to the declarations in scope.

There were 79,183 directions in total, associated with 93,535 declarations and 1,283,445 sets of CP questions and answers.

## 4.3 Data analysis

For each direction, we assessed whether there was a subsequent amendment, where this was defined as an amendment within 30 days. By focusing on this time range, the aim was to restrict the focus to amendments that could be deemed to be in response to the direction.

There were 6207 directions with such amendments. Table 4.1 gives the amendment rates for those directions which occur more than 10 times with an amendment rate of at least 5%, along with the number of each direction.

**Table 4.1:** Broker behavior data: Amendment counts and rates per direction, ranked by amendment rate.

Direction	Rate	Count	Direction	Rate	Count
ICS amendment required - Qtine	76.3%	93	Pending Information	11.5%	348
Hold pending Approval	50.0%	12	Number ICE Container Inspected	10.6%	47
Withdrawn Entry	46.7%	15	Hold Pending Payment	10.4%	364
Administration Withdrawn	32.4%	37	Protocol Fertiliser Inspect	10.4%	48
SIP - Inspect (Unpack)	25.0%	108	Other Approved Method	10.3%	29
FC Audit Release after inspect	23.5%	17	Verify Packing	10.0%	70
Unsupervised	23.5%	17	Tailgate - Rural Destination	9.9%	1195
Administration Staff Comments	22.5%	865	Compliance Agreement Fee	9.6%	3872
Personal effects Inspect	22.2%	27	LCL Inspection	9.3%	787
Plant Product Pathway	22.1%	154	Test seed samples	9.2%	65
Staff Info See Comments Food	20.6%	34	Seals Intact Pending Docs	8.7%	334
ICS Major Amendment (F)	20.1%	612	Exported	8.3%	36
Disinfection	20.0%	10	Seeds Inspect (Sample if req)	8.1%	74
SIP - Document Assessment	19.7%	299	Reordered-in to quarantine	8.1%	963
Check FC Doco	19.0%	116	Officer Hold, contact QTNE	8.0%	299
Mismatched H/O	18.2%	582	Present all documentation	8.0%	18689
Release after Inspection	18.2%	1692	Autoclaving	7.7%	26
SIP - Hold Seals Intact	17.2%	186	Inspect (unpack)	7.6%	4072
Vessel Inspect - Full Inspect	16.9%	89	Inspect (Hold Seals Intact)	7.3%	386
Test and hold	16.8%	2331	Finalised & Released OE Reg	7.1%	28
Break Bulk Inspection	16.7%	401	AEP Random Audit	6.9%	159
AIMS Follow up - See Comments	15.4%	13	CH3Br 48gM3 24hr 21C or above	6.8%	427
Deep Burial	15.0%	20	AIR Freight Inspection	6.4%	358
ICS Major Amendment (Q)	14.0%	1196	Cleaning as Directed	6.1%	66
Automatic release	13.5%	1861	Bulk Timber Inspect (No certs)	5.9%	17
Pending Documentation	13.3%	173	H.T. - 85 C. for 8hrs	5.9%	17
H/O Test and Hold	13.2%	408	Prawn Verification	5.6%	18
Follow Up Inspection Required	12.6%	823	Supervised by Quarantine	5.6%	18
HT 50% Humidity - 95C for 24hr	12.5%	80	Track Commodities	5.3%	301
DAFF Processing Required	12.4%	749	In accordance with permit cond	5.3%	38
Staff Information See Comments	12.3%	244	Under Surveillance	5.0%	20
Broker Non-conformity	11.8%	34			

#### 4.3.1 The impact of threatening amendments

Certain directions were considered “threatening” *a priori*, and likely to prompt an amendment. These were:

- Air Freight Inspection
- ICS amendment required - Food
- ICS amendment required - Qtine
- Inspect (Hold Seals Intact)
- Inspect (unpack)
- Nursery Stock Inspect
- SIP - Hold Seals Intact
- SIP - Inspect (Unpack)
- Under Surveillance
- Verify (Hold Seals Intact)
- Verify certs Bulk Timber Insp.
- Verify Commodity
- Verify Packing
- Verify prior to Man. Fum.
- Verify prior to VolFum.



- Verify Tarping
- Withdrawn Entry

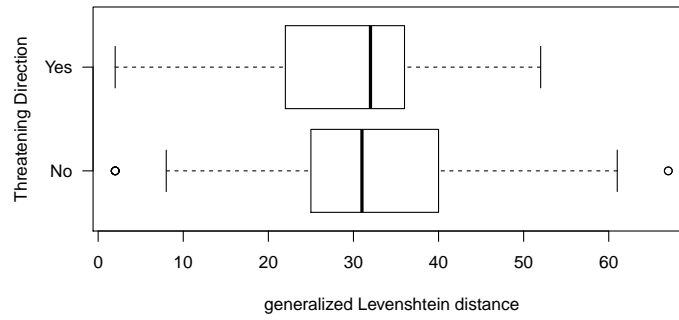
It was therefore of interest to see when these threatening directions resulted in greater amendments, both overall and by specific amendment type. The results for a range of amendment types of interest are given in Table 4.2. These results include odds ratios, comparing the odds of an amendment comparing threatening and non-threatening directions. For example, the odds of any amendment following a threatening direction are 1.12 times higher than the odds of an amendment after a non-threatening direction. We are 95% confident that the true odds is between 1.02 and 1.23; this interval does not include 1 (where the odds would be equivalent) so this might be considered a statistically significant relationship, although the signal is very weak. While there was a significant relationship between threatening directions and overall amendments, there were no specific amendment types that were significantly associated with threatening directions. This is due, in part, to the smaller incidence of amendments of specific types; some of the estimated odds ratios are still reasonably large (or small, depending on the direction of the relationship). Note that there were no amendments to Brokerage ID in the data provided.

**Table 4.2:** Broker record amendment odds ratios after threatening directions, with a 95% confidence interval (CI) and count of each amendment type.

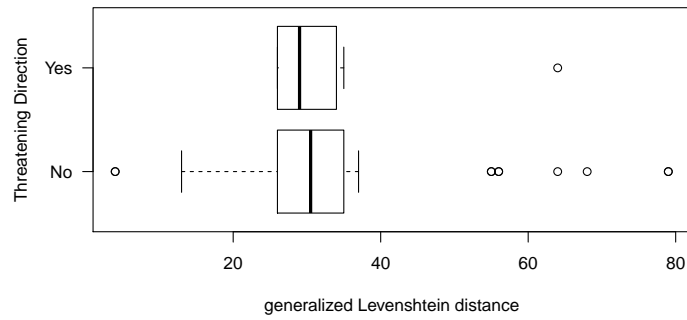
Amendment type	Odds Ratio	95% CI	Count
Any amendment	1.12	(1.02, 1.23)	6207
Change of Importer ID	1.61	(0.73, 3.16)	79
Change of Supplier ID	2.07	(0.39, 7.26)	19
Any reduction in the total number of lines	0.39	(0.05, 1.47)	58
Any increase in the total number of lines	1.18	(0.85, 1.62)	512
Deletion of at least one line number	0.37	(0.04, 1.41)	61
Addition of at least one line number	1.17	(0.84, 1.60)	515
Any change in delivery address	1.24	(0.83, 1.80)	339
Any change in importer address	1.85	(0.84, 3.67)	70
Any change in tariff class	0.61	(0.35, 1.00)	319
Any change in goods description	1.05	(0.78, 1.39)	704
Any change in goods quantity	1.08	(0.77, 1.49)	507

The results in Table 4.2 are for incidence of change, but it may also be of interest to know whether the degree of change impacts the rate of amendments, for those outcomes for which the magnitude of the change can vary. For line numbers, this is the absolute difference in the number of lines; for tariff classes and goods related outcomes, this is the absolute difference in the number of unique examples (as a result of line changes or otherwise); for addresses, a *generalised Levenshtein distance* between the address before and after was calculated. The generalised Levenshtein distance between two strings of characters is defined as the minimum number of changes required to transform one to the other, for which insertions, deletions, and letter swaps all count as a single change. For example the generalised Levenshtein distance between *Tuesday* and *Thursday* is 2; (i) insert *h*, (ii) swap *e* for *r*.

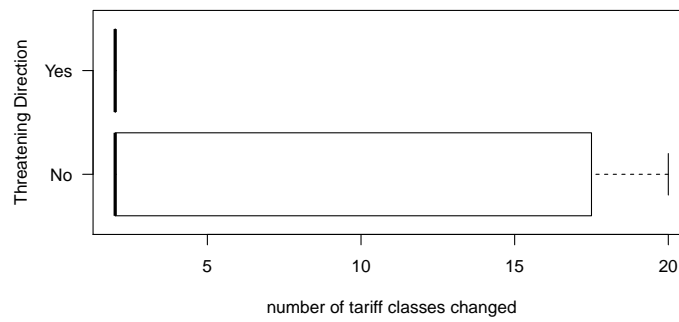
Figures 4.1–4.6 show the magnitude of these different types of changes by threatening direction. Note that these figures exclude those directions with no amendment of this type.



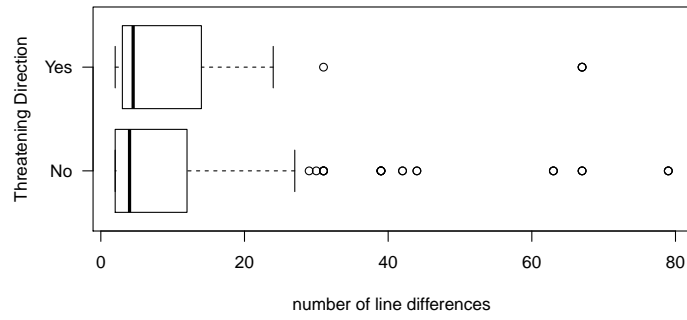
**Figure 4.1:** Magnitude of changes by brokers in delivery address by threatening direction.



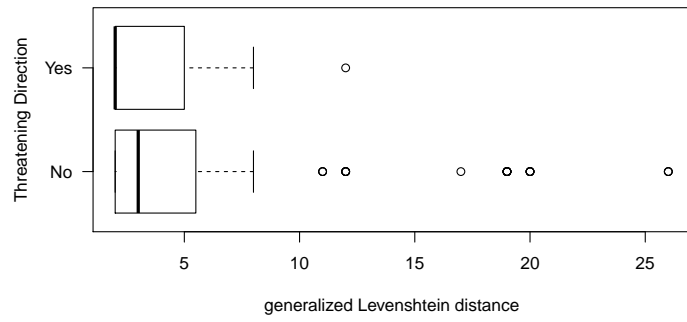
**Figure 4.2:** Magnitude of changes by broker of importer address by threatening direction.



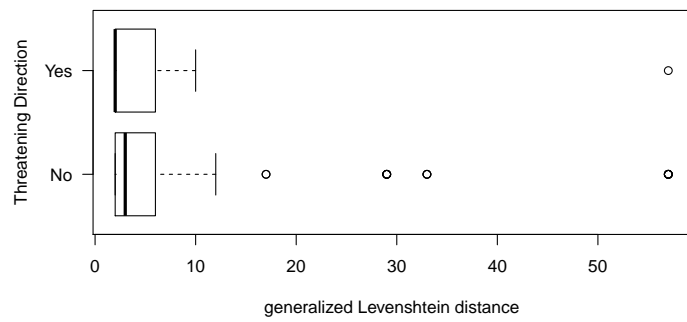
**Figure 4.3:** Absolute difference in number of unique tariff classes by threatening direction.



**Figure 4.4:** Absolute difference in the number of lines by threatening direction.



**Figure 4.5:** Absolute difference in the number of unique quantities by threatening direction.



**Figure 4.6:** Absolute difference in the number of unique goods by threatening direction.

As these figures indicate, threatening directions do not appear to be associated with more substantial amendments; there is, if anything, a tendency towards less serious amendments, particularly for the number of tariff class changes.

### 4.3.2 The impact of special directions

Certain directions were considered “special” *a priori*, and particularly likely to prompt the kinds of amendment that are a concern. These were:

- Inspect (Hold Seals Intact)
- Inspect (unpack)
- SIP - Hold Seals Intact
- SIP - Inspect (Unpack)
- AEP Random Audit
- Follow Up Inspection Required

It was therefore of interest to investigate patterns between these directions and amendments, both overall and by specific amendment type. The results for a range of amendment types of interest are given in Table 4.3.

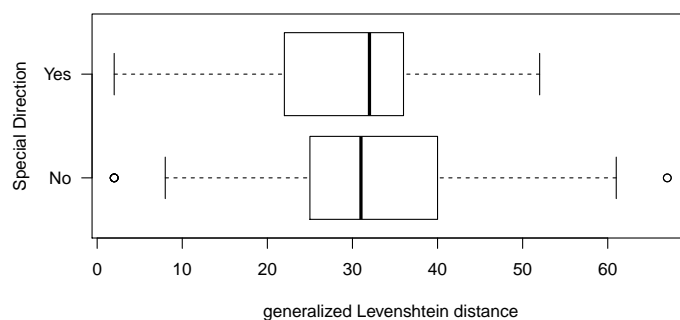
**Table 4.3:** Broker amendment odds ratios after special directions, with a 95% confidence interval (CI) and count of each amendment type.

Amendment type	Odds Ratio	95% CI	Count
Any amendment	1.17	(1.06, 1.28)	6207
Change of Importer ID	1.08	(0.42, 2.36)	79
Change of Supplier ID	2.09	(0.39, 7.34)	19
Any reduction in the total number of lines	0.37	(0.04, 1.42)	58
Any increase in the total number of lines	1.19	(0.85, 1.62)	512
Deletion of at least one line number	0.39	(0.05, 1.50)	61
Addition of at least one line number	1.20	(0.86, 1.64)	515
Any change in delivery address	1.58	(1.09, 2.22)	339
Any change in importer address	1.24	(0.48, 2.73)	70
Any change in tariff class	0.50	(0.27, 0.86)	319
Any change in goods description	0.94	(0.70, 1.26)	704
Any change in goods quantity	0.98	(0.68, 1.37)	507

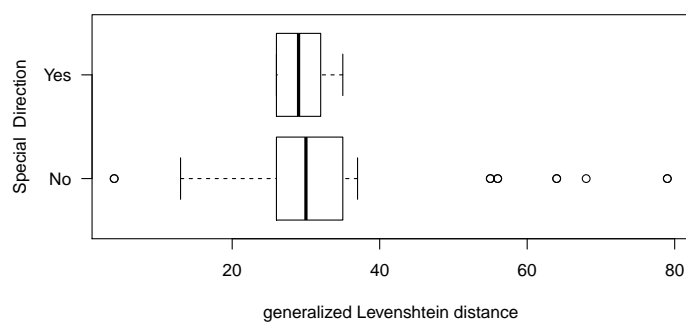
Most of the results are similar as for threatening directions, but there are a few differences. The overall rate of amendment is higher, an increase in overall odds of amendment of 17% for special directions compared to 12% for threatening directions. There is an increase in the rate of delivery address changes (odds of 1.58 compared to 1.24) and a decrease for importer address (1.24 compared with 1.85). The 95% confidence interval for delivery address no longer contains zero, indicating a significant relationship between a special direction and a delivery address change. There is a slight reduction in the rate of amendments of importer ID, though the 95% confidence interval for this remains wide. There is a drop in the rate of tariff class changes for these kinds of amendments; the 95% confidence interval now no longer contains 1, evidence that special directions result in *fewer* tariff class changes when compared to the remaining directions.

As with the threatening directions, results in Table 4.3 are for incidence of change, but it may also be of interest to know whether the degree of change impacts the rate of amendments, for those outcomes for which the magnitude of the change can vary. These are defined in the same way as for threatening directions to produce Figures 4.7–4.12. Note that these figures exclude those directions with no amendment of this type.

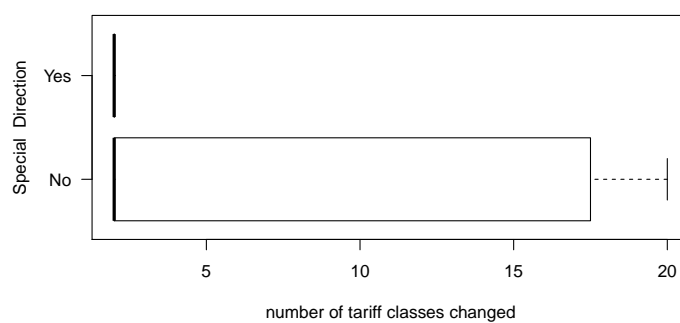
As these figures indicate, special directions demonstrate very similar patterns to threatening directions: no particular difference or a tendency towards less serious amendments, particularly for the number of tariff class changes.



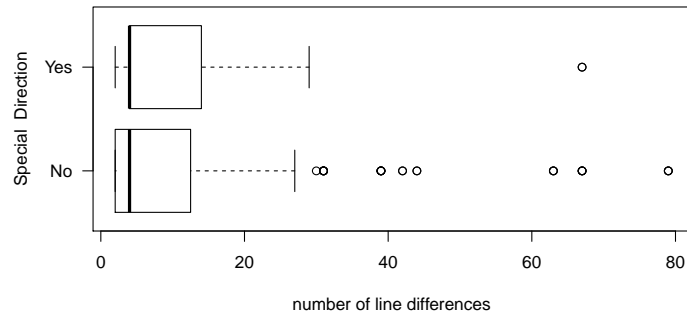
**Figure 4.7:** Magnitude of changes by brokers in delivery address by special direction.



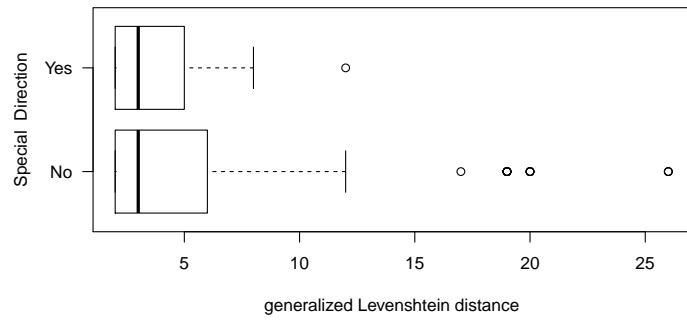
**Figure 4.8:** Magnitude of changes in importer address by special direction.



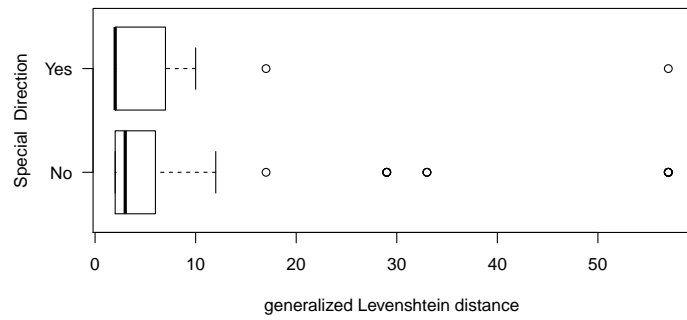
**Figure 4.9:** Absolute difference in number of unique tariff classes by special direction.



**Figure 4.10:** Absolute difference in the number of lines by special direction.



**Figure 4.11:** Absolute difference in the number of unique quantities by special direction.



**Figure 4.12:** Absolute difference in the number of unique goods by special direction.

### 4.3.3 Community Protection questions and amendments

Answers to key Community Protection (CP) questions for each amendment were also available, and changes in these answers as part of amendments is of interest. Certain answers are considered “high risk” and are the most relevant from the point of view of quarantine risk.

For each amendment that occurred in response to a direction, the difference in the incidence of any and the rate of all high risk answers before and after could be compared. Table 4.4 gives the incidence of any high risk answer before and after amendments. As this table indicates, very few amendments resulted in the introduction of a high risk answer but a quarter resulted in the removal of any high risk answers.

**Table 4.4:** Incidence of any high risk answer from the broker before and after amendment with row percentages.

		After amendment	
		no high risk answer	high risk answer
Before amendment	no high risk answer	325 (96.7%)	11 (3.3%)
	high risk answer	1517 (25.8%)	4354 (74.2%)

There is also a general reduction of the rate of high-risk answers across the lines. Of the 450 amendments that resulted in a change in the proportion of high risk answers, in 289 of these (64.2%) the change was a reduction in the rate of high risk answers, with the median reduction rate being 0.10 or 10%.

The nature of the 1517 amendments that accompany the removal of a high risk answer has been explored in Table 4.5. An odds ratio greater than one suggests that the amendments of this nature tend to result in CP high-risk reduction, whereas an odds ratio less than one suggests amendments of this nature tend to result in CP high risk increase. There are many consistent results with the general amendment patterns in relation to the threatening and special directions. The same amendment types that tend to be associated with threatening and special directions are also those associated with CP high risk status removal. The results in Table 4.5 are generally stronger with few 95% confidence intervals containing 1. The only exception to this is tariff class, where the results were reversed. While there was evidence of fewer tariff class amendments after threatening and special directions, these tariff class amendments were highly associated with CP high risk removal, as were importer ID and address amendments, increases in line numbers, and changes to goods quantity.

**Table 4.5:** Broker amendment odds ratio for CP high risk removal, with a 95% confidence interval (CI) and percentage of CP high risk reductions that include that type of amendment.

Amendment type	Odds Ratio	95% CI	% of CP high risk reductions
Change of Importer ID	2.13	(1.31, 3.42)	2.1%
Change of Supplier ID	0.82	(0.20, 2.59)	0.3%
Any reduction in the total number of lines	0.23	(0.06, 0.62)	11.9%
Any increase in the total number of lines	1.78	(1.46, 2.17)	0.3%
Deletion of at least one line number	0.40	(0.15, 0.88)	12.1%
Addition of at least one line number	1.82	(1.49, 2.21)	0.5%
Any change in delivery address	0.62	(0.46, 0.84)	2.8%
Any change in importer address	2.63	(1.59, 4.35)	2.1%
Any change in tariff class	2.17	(1.71, 2.75)	8.4%
Any change in goods description	1.03	(0.86, 1.24)	11.6%
Any change in goods quantity	1.36	(1.11, 1.67)	10.0%

## 4.4 Conclusion and Recommendations

This report combines import declarations and quarantine directions to assess the amendment patterns in response to directions. The analysis has been performed on a subset of ICS data. There is little evidence of relationships between threatening directions and specific amendment types. There is some reasonably

strong evidence of associations between certain amendment types and the removal of a Community Protection high risk answer.

**Recommendation 9.** The department should note that there is evidence in the broker analysis case study to suggest that some brokers are altering CP status within declarations to reduce the apparent biosecurity risk after receiving a quarantine direction. Further analysis of this phenomenon may provide actionable intelligence.

This project did not involve physical inspection of the goods associated with the record that was amended by the broker, so the true biosecurity risk is unknown. There may be some benefit to targeting such records as part of the Cargo Compliance Verification exercise (see Chapter 7).

**Recommendation 10.** The broker analysis project did not involve the physical inspection of consignments for which tariff codes had been changed, which represents a possible avoidance behaviour by brokers. The department should consider expanding cargo surveillance to include those AIMS entries that are modified to appear *less risky* by the broker upon receipt of any quarantine directions.

Finally, the first point at which brokers may game the regulatory system is absent from this study, namely when they provide a low risk answer to a community protection profile question, and the goods are not referred to the department. There may be some benefit to targeting such records as part of the Cargo Compliance Verification exercise (see Chapter 7).

**Recommendation 11.** The department should note that the broker analysis project did not involve examination of goods that were not referred to the department because of provision of a low-risk answer to community protection profile questions. There may be some benefit to targeting such records as part of the Cargo Compliance Verification exercise (see Chapter 7)



## Chapter 5

# Risk Factor Prediction for International Vessels

SANDY CLARKE\*, AND NIANJUN LIU†

### 5.1 Summary

This chapter<sup>1</sup> reports a data mining exercise to identify risk factors using data from the Vessel Management System (VMS) dataset, in order to be able to predict inspection failure. This chapter has subsequently been published as Clarke et al. (2017).

#### 5.1.1 Background

International vessels are known to present biosecurity risk. The principal means are from biofouling, in which case pests are attached to a surface of the vessel and disembark in Australian waters, and the exchange of ballast water, in which case pests are captured during the intake of ballast water and subsequently discharged when the tanks are partially or fully emptied.

#### 5.1.2 Motivating Question

Are certain kinds of vessels more likely to be carriers of biosecurity risk than other kinds? Can we predict the outcome of a first-port inspection based on readily available information about the vessel?

#### 5.1.3 Methods

The DAWR vessel inspection data from 1 July 2006 to 31 October 2013 were available for analysis. The database contained a large source of raw operational data including information on voyages, visits and inspections, and included vessel name, vessel type, call sign and International Maritime Organisation (IMO) number, agent name, visit date and time.

Classification trees with gradient boosting were used to assess characteristics that predict high-risk vessels ( $n = 93,006$ ) for carrying BRM, across the seven years of inspection data.

#### 5.1.4 Results

Undeclared vessels and suspected irregular entry vessels posed the highest risk, but both were rare. Vessels that visit infrequently ( $\leq 20$  visits in 7 years) were common and had almost three times greater odds of failing inspection than vessels visiting frequently. On statistical analysis, yachts appeared to pose less risk than commercial vessels. In operational terms, a tentative profiled 20% fraction would contain 57% of genuine failures, and the concomitant non-screened group would contain 82% of passes. The most

---

\*Statistical Consulting Centre, The University of Melbourne

†Research and Intelligence, Biosecurity Implementation Branch, Department of Agriculture and Water Resources.

<sup>1</sup>This chapter is here reproduced from Clarke et al. (2015a), for completeness of this report. Recommendations and the summary section are added.

common reason for inspection failures was ballast water non-compliance (2.53%) and plant or insect detections (1.77%); biofouling was less common (0.13%) but testing for biofouling is not exhaustive.

### 5.1.5 Conclusions and Future Directions

We propose predictive models that have potential for implementation in future screening. All the R scripts required to produce these models have been provided to enable replication and adaptation.

## 5.2 Data preparation

The data provided for this sub-project were vessel inspection data from 1 July 2006 to 31 Oct 2013. Of these, only **routine** inspections with a result are relevant, although other inspection types and results were included in the database. The key outcome is whether the inspection result was a failure/non-conformity, as opposed to a pass. A pass means that the vessel is relieved of quarantine obligations, a failure means that some treatment or corrective action must be undertaken before the vessel can be released, and non-conformity is an intermediate outcome in which the inspector considers an outcome to be non-compliant but not disastrously so. Both failure and non-conformity will be referred to as a failure for the purposes of this sub-project; because both constitute relevant non-compliance. Using this definition, the data set contains 82,991 passes and 10,015 failures for the purposes of analysis.

A separate file with 39,486 entries containing the names of shipping agents was supplied, because agent was identified as a likely risk factor by Seaport experts. Using inspection date, time and imo.number, matches could be established for 23,775 of the inspections. The remaining inspections were coded as ‘unmatched’ for agent name.

The fields available for modeling are listed in Table 5.1. Note that all of these fields can be treated as categorical variables.

**Table 5.1:** Potential factors currently available for modelling biosecurity risk in vessels.

Field Name	Field Definition
agent.name	Shipping agent name
current.port	Port of a vessel mooring at the current time of recording
inspect.month	Month of an inspection done
inspect.quarter	Quarter of an inspection done
inspect.year	Year of an inspection done
last.country	Last country of a vessel visiting
last.port	last port of a vessel visiting
pdv.voyage	Whether a voyage is qualified for PDC clearance
pdvchangereason	Reason why the status of the vessel PDC was changed
pdccycle	Whether it is in a pdccycle?
pdvvisitcount	Number of visits with a qualified PDC clearance
pratique.visit	Whether a visit is pratique or not?
proclaimedport	Whether it is a proclaimed port?
regioncode	Code of region
vessel.name	Name of the vessel
vessel.type	Type of the vessel
visit.is.first	Whether it is the first visit during the whole voyage?
visit.is.last	Whether it is the first visit during the whole voyage?
visit.length.category	Whether the vessel is commercial or yacht?
visit.month	Month of a visit
visit.quarter	Quarter of a visit
visit.ship.type	Whether it is a ship or yacht?
visit.vessel.class	Selected from the classes of “Government, Cruise vessel, Commercial, livestock vessel, Yacht, Undeclared”
visit.year	Year of a vessel visiting a port

PDC refers to pratique documentary clearance, which is a risk-based algorithm used by the department to reduce intervention effort systematically for low-risk vessels with a proven history of compliance.

Other fields available in the data file were excluded because they had only one level or because they were found to be a feature of the failure itself, rather than a potential risk factor for detecting failures. Even though the visit and inspection dates were similar, the presence of differences necessitated consideration of both. The department has confirmed that Table 5.1 contains all the risk factors that are considered potentially important at the time of publication of this report.

The first data cleaning step required was to collapse those categorical fields that had many levels, including some with very small counts, as follows. For last country, current port and agent name, levels with less than 100 counts were collapsed into two groups: ‘lowest’ (<20 counts) and ‘lower’ (20-99 counts). For last port and vessel name, there was a particularly large spread of possible levels, so levels with less than 200 counts were collapsed into three groups: ‘lowest’ (<20 counts), ‘lower’ (20-99 counts) and ‘low’ (100-199 counts). As well as simplifying these fields for analysis, this enables future prediction on any new, rare levels to be modelled, treating them as if they were in one of these categories.

There were very few missing values for these data, aside from agent name (about 75% missing due to the failure to match) and pdccyle. The missing values were included as a level for the purposes of modelling. Visit length category had a level labelled ‘Unknown’ and vessel class had a level ‘Undeclared’, both of which were included as levels for the purposes of modelling.

For the purposes of analysis, all fields were converted into a series of binary variables, one for each level (1 indicating that level was present for the given record and 0 otherwise), to reduce the data complexity and separate the effects of individual levels more clearly. This resulted in 348 binary candidate predictor variables in total.

## 5.3 Data analysis

An initial reduction of fields was performed before a complex model with all fields and their interactions could be considered. In order to do this reduction, the dataset was randomly divided into two halves, a training set and a test set. The training set was used to fit a simple generalised linear model which was then used to predict the outcome for the test set. The performance of this was assessed using the area under the ROC curve (AUC) for the test set; a common means for assessing predictive performance (see Section 3.3 for a brief description of ROC and AUC). This process was repeated 5 times and summarised by the average AUC. Any field which had an average AUC greater than 0.505 was carried over into the next analysis step; note that an AUC of 0.5 or less indicates no predictive power. Fields were considered independently of each other at this stage.

Table 5.2 contains the list of the 63 fields and levels that were carried into the next phase of analysis, because of evidence of individual predictive power. Note that predictive power doesn’t imply only greater risk; this also includes fields which show notably lower risk.

There are a few things to note at this stage. The ‘lower’ and ‘lowest’ categories of vessel names (constructed as part of the data cleaning process described above) feature in this table, indicating that vessels which occur rarely in the data set seem to be related to risk. Also, the featured years, quarters and months are the same for both inspection and visit dates (e.g., the inspection month of June and the visit date of June both feature), unsurprising given that these dates are very closely related. Although agent name was anticipated by the department staff to be a key predictor, none of the agent names has demonstrated particular predictive power. Of course, a large number of the inspections did not have a match for the agent name, limiting the potential for this predictor.

Once the set of predictor fields was sufficiently reduced, a model combining these fields could be used. The model approach used involved decision trees with boosting and shrinkage, using the gradient boosting machine package (`gbm`) in R (R Core Team, 2013). More details of this approach can be found in Appendix A. Tree-based models progressively split observations into two groups, where the splits are based on levels of the predictors. This approach allows for complex relationships between predictors and outcomes, including interactions between predictors in the way they impact the outcome, and it can handle a range of different predictor types as well as predictors with missing values. The boosting and shrinkage aspect involves fitting these tree models many times, for each new step, weighting those observations that were hardest to classify correctly in the previous step.

This is a reasonably computationally intense method, but can still be implemented on a standard desktop computer with the initial data reduction steps discussed above. Should further computational reduction be required, the `randomForest` package within R might be a useful alternative. This involves

**Table 5.2:** Data fields for biosecurity risk analysis after initial feature selection in vessel inspection data.

Field name	Levels
vessel.type	BULK.CARRIER CONTAINER.VESSEL CRUISE.VESSEL GENERAL.CARGO IRREGULAR.FOREIGN.FISHING.VESSEL SUSPECT.IRREGULAR.ENTRY.VESSEL TANKER YACHTS
inspect.month	6
inspect.quarter	Q2 Q4
inspect.year	2007 2010 2011 2012 2013
current.port	ADELAIDE BRISBANE BUNDABERG CHRISTMAS.ISLAND DALRYMPLE.BAY DAMPIER FREMANTLE GLADSTONE HAY.POINT MELBOURNE NEWCASTLE PORT.HEDLAND PORT.WALCOTT SYDNEY TOWNSVILLE
last.port	lowest NOUMEA SURABAYA.JAVA
last.country	CHINA INDONESIA JAPAN NEW.CALEDONIA NEW.ZEALAND PHILIPPINES VANUATU
regioncode	N NE SE SW
vessel.name	lower lowest
visit.is.last	TRUE
visit.month	6
visit.quarter	2 4
visit.year	2007 2010 2011 2012 2013
visit.length.category	Unknown Yacht
visit.vessel.class	Cruise.vessel Undeclared Yachts

the averaging of many trees, generated from random samples of the data, and provides results in a very similar format.

The methods used here draw on Miller et al. (2009)<sup>2</sup>.

## 5.4 Results

As this approach involves many tree models, there is no single tree or diagram that can be used to display the results. However, it is possible to consider the effect of each predictor, averaged over the other fields, using overall measures of importance.

Table 5.3 provides a list of these relative measures of importance, ordered by importance, for the first 18 field/level combinations. These importance values are only relative and have been standardised so that the largest value is set to 100. Each row also contains an odds ratio for the relative odds of non-compliance for passengers with that variable level compared to those without it. For example, if the vessel class is ‘Undeclared’, the overall odds ratio is estimated to be 26.7; that is, the odds of failure if the vessel class is undeclared are more than 25 times higher than other vessel classes. Even though we have increased the relative number of non-compliants in the sample used to generate the model, this is a consistent estimate of the odds ratio for the population, as this is a relative measure. Increasing the number of non-compliant records will bias the estimate of the intercept, but not of the odds ratios. The final column of Table 5.3 gives the rarity of this field level as a percentage. While the vessel class being undeclared is a strong risk factor, this only occurs in 0.4% of vessels.

This does not provide an indication of the nature of any interactions between fields in the model, though undoubtedly such effects are going to be involved. Such effects are included in the model itself and in any prediction arising from the model.

These results show that vessels with fewer than 20 visits are more likely to fail biosecurity inspection than vessels with more visits, that yachts are far less likely to fail than other vessel types, and that of the remainder of vessel types, bulk carriers are the most likely to fail. Here and elsewhere the results reported are marginal, as opposed to joint (meaning that they report the effects one-by-one, rather than all at once)..

<sup>2</sup>Available at <http://jmlr.org/proceedings/papers/v7/miller09/miller09.pdf>

**Table 5.3:** Relative importance, crude risks, odds ratio and rarity, for top 18 field levels for modeling biosecurity risk from vessel inspection data.

Field and level	Importance	Odds Ratio	Rarity
visit.vessel.class Undeclared	100	26.7	0.4%
vessel.name lowest	63	2.2	65.1%
regioncode NE	63	1.54	29.0%
vessel.type Not YACHTS	44	16.7	94.3%
regioncode N	40	1.62	12.7%
vessel.type BULK.CARRIER	37	1.6	54.3%
current.port CHRISTMAS.ISLAND	33	4.16	1.3%
visit.year 2012	29	1.2	13.6%
vessel.type SUSPECT.IRREGULAR.ENTRY	26	328.9	0.2%
current.port PORT.HEDLAND	22	1.2	7.7%
current.port ADELAIDE	22	2.99	0.8%
current.port DALRYMPLE.BAY	22	2.53	3.4%
inspect.year 2013	22	1.49	11.0%
inspect.year 2012	20	1.2	13.6%
current.port Not BRISBANE	19	1.75	90.0%
visit.length.category Not Yacht	19	4.17	93.9%
last.country Not JAPAN	18	1.54	83.2%
current.port HAY.POINT	17	2.65	2.4%

Although the full data were used to determine the risk factors, in order to best assess the performance of the model in terms of prediction, it is appropriate to separate the data into testing and training data.

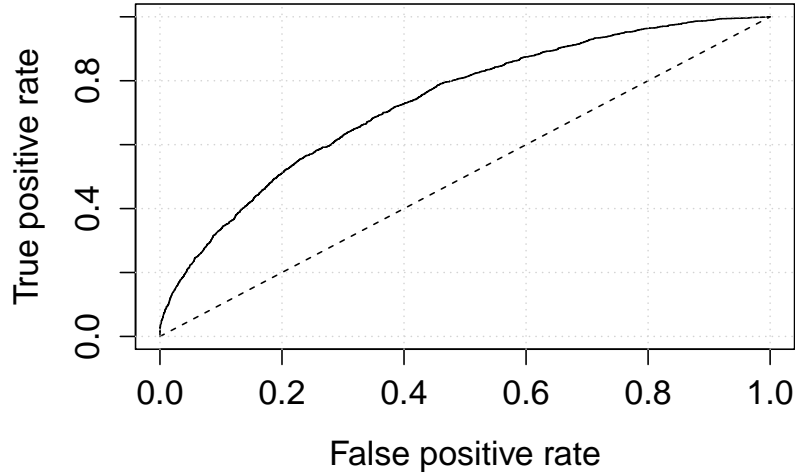
We used the same modelling as above for a random 80% of the data, followed by an assessment of the performance of this model on the 20% not used for the model. This assessment involved nominating a screening proportion and then selecting that proportion of the highest-risk vessels, using risk as predicted by the model. We then assessed the failure rate within that sample (the sensitivity) and the fraction of the total passing vessels outside the sample (the specificity). We want both of these quantities to be high, but there is typically a trade-off between them; if we screen more we will detect more failures, but we will also spend resources screening unnecessarily.

Table 5.4 gives the results for a range of sampling fractions. For example, if we take the top 20% in terms of risk as predicted by our model, and compare these to whether they actually failed, we find that our screened group contains 45.8% of the genuine failures (929 out of 2027) and our non-screened group contains 83.2% of the passes (13728 out of a total of 16506 passes). This corresponds to an odds ratio (OR) of 4.2 (95% CI: 3.8, 4.6). As we increase the screened fraction, the sensitivity improves (we find more failures) but at the expense of specificity (we also increase our false positives).

We can also represent this performance on a continuum, using the receiver operating characteristic (ROC) curve in Figure 5.1. The dashed line depicts random sampling with no profiling; if the statistical models were no better than random the we would expect to see the curve close to this line.

**Table 5.4:** Performance of the vessel inspection prediction model for biosecurity risk on test data, for a variety of sampling fractions.

Fraction	Sensitivity	Specificity	OR	95% CI
10%	28.7%	92.3%	4.82	(4.30, 5.40)
20%	45.8%	83.2%	4.18	(3.79, 4.61)
30%	58.9%	73.5%	3.98	(3.61, 4.38)
40%	69.5%	63.6%	3.98	(3.60, 4.41)
50%	79.3%	53.6%	4.43	(3.96, 4.97)
60%	85.6%	43.1%	4.51	(3.97, 5.14)
70%	91.0%	32.6%	4.90	(4.19, 5.76)
80%	95.7%	21.9%	6.26	(5.03, 7.88)
90%	98.7%	11.1%	9.59	(6.50, 14.77)



**Figure 5.1:** The performance of the vessel risk prediction model on test data, depicted as a ROC curve.

The choice of a screening fraction will depend on the cost of screening relative to the seriousness of failing to detect a genuine failure.

## 5.5 Conclusion and Recommendations

The aim of this sub-project was primarily to determine the key risk factors for failures in the VMS data set. However, the kinds of results presented, combined with the scripts made available, also enable ongoing predictive modelling based on the risk factors identified. It is also possible to update these models as new data become available and the risk factors change.

The decision of whether or not to use the risk factors identified in this chapter as a screening tool depends on several factors, most particularly the relative cost of inspecting vessels compared with the cost of missing non-compliance. If profiling were performed upon international vessel arrivals using the strongest signals arising from the model exercise, namely the influence of the number of visits and the vessel type, then the benefits may not differ substantially from the current PDC policy.

**Recommendation 12.** For marine vessels, the identity of the agent was considered *a priori* to be a field that might best identify vessels with better or worse governance, and therefore possibly better or worse biosecurity compliance. The department should improve the quality of agent information for international marine vessels so that more inspection outcomes can be linked to the vessel agent, to provide a more rigorous test of this conjecture.

## Chapter 6

# Overview of Transfer Learning

SANDY CLARKE\*

### 6.1 Summary

This chapter<sup>1</sup> is designed to provide an overview of the area of transfer learning. No further work on this study was undertaken after submission of the current chapter because of time and resource constraints.

#### 6.1.1 Background

Transfer learning is what human learners do inherently but is relatively new in the area of machine learning. The idea behind transfer learning is that related information is introduced to aid in a given task, beyond the data that would typically be used, that related directly to that task. The success of this transference depends heavily on the genuine relatedness of the additional data, but can be used to great effect when suitable auxiliary data are present.

#### 6.1.2 Motivating Question

Can transfer learning be used to develop information about biosecurity risk management, specifically information that is otherwise expensive or impossible to obtain, based on information that is cheap and easy to obtain?

#### 6.1.3 Methods

We provide a high-level review of transfer learning techniques and examples of applications.

#### 6.1.4 Results

No results arise.

#### 6.1.5 Conclusions and Future Directions

No specific recommendations are made for this study. Transfer learning may be of use to the department, especially in the area of estimating large numbers of related risks.

### 6.2 Relevance

All risk assessments, including those in the department's Risk Return Resource Allocation (RRRA) model, require estimates of a large number of parameters, including the rate of arrival and release of biosecurity risk material into Australia and the behaviour of individual pest and diseases. Direct estimates are usually unavailable due to complete lack of data, incomplete datasets or unbalanced occurrences within departmental databases.

---

\*Statistical Consulting Centre, The University of Melbourne

<sup>1</sup>This chapter is reproduced almost verbatim from Clarke et al. (2014) for completeness.

Transfer learning is a method to estimate values for which there is scarce or incomplete evidence. The basic idea is that when there is insufficient data to directly support the evaluation of a specific value, the value can be estimated and transferred from other related data sources by assessing the similarity and dissimilarity among them. This study will provide guidelines to identify situations in which transfer learning might be applied and where it cannot.

For example, the performance indicator case study undertaken by the Border Compliance and the Australian Centre of Excellence for Risk Analysis (ACERA, the predecessor of CEBRA) used empirical Bayes methods to ‘share’ information about estimates of leakage rate among different passenger cohorts. In this way, information from large, well-known cohorts was pooled with information from small, less well-known cohorts to improve the performance of the profiles. This sub-project will determine whether empirical Bayes methods, for example, can be used for other pathways.

## 6.3 Terminology

The first step in understanding the area of transfer learning is to learn the language, which reflects the fact that this area has arisen out of engineering and computer science rather than statistics.

The terminology draws on the terminology used in machine learning more generally, within which each row or source of data is referred to as an instance, each of which has corresponding features, ordinarily arranged in columns. In the context of border compliance data, an instance might be a given passenger record, with the features being the fields available for each record.

In supervised machine learning, there are also labels corresponding to the outcomes, some of which are assumed to be known (for training data) and some of which are not (for test data). A label could be whether or not a passenger was found to have biosecurity risk material, which would be known for a set of training data but unknown for some future test data. The most common classification and regression problems fall within this framework, with the aim of determining which features best predict the labels. For example, a question of interest might be which passenger data fields relate most strongly to the presence or absence of biosecurity risk material on the passenger.

Classification when no labels are available is referred to as unsupervised machine learning, and usually involves clustering or dimension reduction of the feature space.

### 6.3.1 Formal transfer learning terminology

The most systematic transfer learning notation is that given by Pan and Yang (2010) and has been presented below, although drawing partly on terminology used elsewhere.

The two key machine learning definitions used by Pan and Yang (2010) are a domain and a task. A domain includes the features,  $X$ , along with their probability distribution,  $P(X)$ . Given a specific domain, a task refers to the assignment of a label,  $Y$ , and an objective prediction function,  $f()$ , which is typically of the form  $P(y|x)$ .

For example, in the standard regression framework, the domain is the predictors or explanatory variables and the task is the estimation of the outcome or response variable. The aim in this case is also to establish the best function,  $f()$ , that relates the predictors and the response for the purposes of understanding the relationship or for prediction of  $y$  from  $X$ .

Transfer learning assumes there are two types of domains and tasks, namely the source and target types. The source data are typically better understood but not of particular interest, whereas the target predictive function,  $f_T()$ , is the goal of the exercise, and seeks to benefit from the comparably better information available about the source.

Transfer learning assumes that the source and target domains and/or task differ, otherwise this would be a traditional machine learning problem. However, it may be that either the domains or the tasks are the same. There are two broad characterisations of transfer learning, namely inductive transfer learning and transductive transfer learning.

*Inductive transfer learning* refers to the case where the source and target tasks are different. For the prediction process to be possible, there must be some labelled data in the target domain, but there need not be any in the source domain. It is considered induction because the source data is used to understand a process which is generalised and applied to the target case.

*Transductive transfer learning* refers to the case where the source and target tasks are the same but the source and target domains are different. There is no labelled data in the task domain but there must be some labelled data in the source domain. It is considered transduction because the training process is based on the same tasks as the testing process.



As with regular machine learning, it is also possible to perform unsupervised transfer learning for cases where there are no labelled data but clustering or dimension reduction is required. This approach is unlikely to be relevant for the current data mining project, but is worth noting nonetheless.

Of course, when the source and target domain and tasks are the same, the two data sources can be amalgamated for the purposes of standard machine learning; the source/target distinction is irrelevant.

These processes can be data-driven or human-driven. That is, prior knowledge from individuals may guide the formulation of the relationships, or shared parameters and tasks. As with all machine learning, the desire is for increased automation but the nature of the exercise often still requires human intervention to select relationships that are likely to increase learning.

For example, if a data set were available from a border compliance division of another country, the potential gain in incorporating such data may need to be assessed by those who understand the similarities and differences between the procedures used in that country. Data-based approaches may also be available to make such assessments, such as assessing whether the patterns in the fields and their relationship to non-compliance are consistent.

The precise methods used in transfer learning rely heavily on existing machine learning algorithms. In particular, transfer learning can be seen as an extension to multitask learning which aims to simultaneously predict multiple tasks. The difference is that transfer learning is concerned with performance on one task while multitask learning gives the same weight to different tasks. Transfer learning also doesn't require as much data because the task is focused; for example, non-target data do not require labels.

## 6.4 How to transfer

There are four main classes of approaches to transfer learning, *instance transfer*, *feature representation transfer*, *parameter transfer* and *relational knowledge transfer*.

Parameter transfer and relational knowledge transfer require commonality between the source and target domains, and hence are only possible within the inductive transfer framework.

### 6.4.1 Instance transfer

Instance transfer assumes that the information in the source domain can be reweighted to be used for learning in the target domain. It is analogous to the analysis of data from complex sample surveys.

Methods used to perform this type of transfer typically focus on cases where the source and target data are the same, or at least overlap, but the distributions differ. Therefore estimating the differences in the distributions becomes essential. These methods often involve iterative reweighting such as in Adaptive Boosting and Support Vector Machines, as well as importance sampling. Literature in the area of sampling bias is relevant to this type of transfer.

### 6.4.2 Feature representation transfer

Feature representation transfer assumes that a particular feature representation can improve the similarity between the source and target domains. This representation can be developed on unlabelled data and then applied to labelled data for the purpose of performing the target task.

Methods used to perform this type of transfer focus on sparse learning methods, seeking a low-dimensional representation that is common across tasks, analogous to process of dimension reduction before performing regression.

### 6.4.3 Parameter transfer

Parameter transfer assumes that the source and target tasks share some parameters which can be used to code the transfer. This is the approach most directly related to multitask learning which commonly involves shared parameters across tasks. The numerous methods designed for multitask learning can be used in this case including Support Vector Machines and hierarchical Bayesian frameworks.

The closest analogy to this approach is path analysis, which can be used to connect a set of variables beyond the standard 'one path' regression approach.

#### 6.4.4 Relational knowledge transfer

Relational knowledge transfer assumes that there are some relationships between the data within the source and target domains that are common. This is different from parameter transfer which requires parameters in common; here we exploit common or analogous relationships within each domain. An example of an approach which utilises this is Markov Logic Networks. This is the least common of the four main transfer types, relying on situations where we want to predict links between instances rather than class membership of instances.

#### 6.4.5 Available resources

The platform of choice in the area of transfer learning is Matlab with the vast majority of open source software relying on this platform<sup>2</sup>. There are a range of R packages that implement machine learning methods which could be used for such analyses but none specifically dedicated to transfer learning alone (Hothorn, 2013).

### 6.5 When to transfer

Much of the literature on transfer learning is concerned with the processes of performing analyses (the 'how') but it is important to ask whether it is appropriate to transfer in the first place. In particular, we must know how to identify when transfer learning is likely to introduce such bias that it will hurt our predictive performance, also known as *negative transfer*.

It comes down to how well related the source and task domains and/or labels are as to whether this process will be effective. Rosenstein et al. (2005) and Torrey and Shavlik (2009) explore negative transfer in greater detail, recognising that the opportunity for gain and the risk of negative transfer are traded off in different methods. In the same way that training data can be used to assess other machine learning algorithms, it is often possible to compare results with and without transfer to determine if there is sufficient evidence of gain.

### 6.6 Examples

It is the nature of such a broad area that it is not possible for all methods to be captured with a small number of examples. However, it is helpful to demonstrate how these concepts might be put into practice, in the framework of applications to image recognition, an area where these methods are popular.

#### 6.6.1 Image recognition

Many of the textbook examples of transfer learning found in the literature relate to the process of image recognition (e.g., Argyriou et al. (2008); Lim (2012); Raina et al. (2007)). Typically, the task is to identify whether a particular object is present in an image, say, a letter or number. Each image is an instance and each image that is known to contain or not contain the object is labelled. The target task is to identify which images contain the object from a set of unknown images. The data itself are usually pixel level measurements. Similar applications include speaker identification from audio files and the development of email spam filters.

Where a sufficient amount of labelled (training) images are available and no additional (source) information, this is a standard classification exercise. However, where the existing training process is inadequate and source information is available, improvements in performance can be made. The internet yields an abundance of source images, though these are rarely labelled.

In the case of image recognition, we can imagine a range of kinds of additional source information and how it can be used. Here are just a few examples:

- We could borrow labelled training examples from related image sources. We would need to take care in determining whether they are sufficiently related and may wish to weight these differently to the target training images. This could be considered an example of instance transfer.
- We may have obtained our images from the web with text tags but also obtained a number of unrelated text documents from the web. The relationships between words in the text documents could help to classify images. This could be considered an example of parameter transfer.

---

<sup>2</sup><http://www.cse.ust.hk/TL/>

- Given the likely access to a vast amount of unlabelled source data from the internet, we could seek to characterise the feature representation of these images. For example, we may find that it is more powerful to represent the images based on the position and nature of edges rather than pixels. This could be considered an example of feature representation transfer.
- We may expect that the orientation of the object will matter and therefore use source images with a range of (known) orientations of other objects to train our models to perform well at the task of identifying orientation. The results of this can then be utilised for the target task. This could be considered an example of parameter transfer.
- We may have access to source images that contain some features of the target object. We could then use co-clustering to uncover a better data representation to benefit the target set. This could be considered an example of feature representation transfer.

None of the examples above was an example of relational knowledge transfer. Most examples used to illustrate this point in the literature have to do with movie databases, relating movies to actors or directors, etc., rather than problems in image recognition, which provide a rich source of examples for the other approaches.

### 6.6.2 Border compliance

Given our application to border compliance, it may be helpful to also provide some examples in this context.

Take the case where our task is to identify whether biosecurity risk material is present on a passenger arriving to Melbourne Airport from an international airport. Each passenger is an instance and some passengers have been screened for biosecurity risk material. The features are the fields available for each passenger. Transfer learning will be possible when other, supplementary data are available, say, from other contexts or with different fields. Here are a few examples:

- We could borrow screening data from another a secondary airport. We would need to take care in determining whether they are sufficient related and may wish to weight these differently to the records from the airport(s) of interest. This could be considered an example of instance transfer.
- We may have some other passenger screening data for which a different but overlapping set of fields are available. The relationships between these extra fields could help us to better understand passenger behaviour, or compensate for missing or empty fields. This could be considered an example of parameter transfer.
- Given the likely access to a vast amount of historical passenger data for which no screening is available, we could seek to characterise the relationships between fields for these records. For example, we may find that some fields are highly correlated and better represented in some kind of amalgamated way. This representation could then be incorporated into the modelling of records for which screening data are available. This could be considered an example of feature representation transfer.

## 6.7 Conclusion and Recommendations

This study was discontinued because of time and resource constraints. No specific recommendations are made for this study at this time.

## Chapter 7

# Performance Indicators for Cargo Compliance Verification

SANDY CLARKE\* AND JOSE ARIAS<sup>†</sup>

### 7.1 Summary

This chapter<sup>1</sup> proposes some appropriate means of reporting the compliance across some cargo pathways for the Performance Indicators for Cargo Compliance Verification (CCV) sub-project.

#### 7.1.1 Background

The following description is excerpted from the Department’s website.<sup>2</sup> Cargo Compliance Verification (CCV) inspections are conducted by the Department of Agriculture and Water Resources on containerised sea cargo imported into Australia. This is a part of the system of verification that ensures the integrity of the Australian biosecurity system.

A robust biosecurity system helps protect Australia’s agricultural industries, economy, human health and environment from exotic pests and diseases. The department, along with other jurisdictions, industry and the community, plays a vital role in maintaining Australia’s enviable biosecurity status.

Through the inspection program the Department aims to ensure that:

- Importers and their agents comply with all import requirements;
- The department’s intervention program is operating effectively and targeting those imports that pose the greatest biosecurity risk; and
- Information on emerging biosecurity risks, including from commodities which are not typically directed for inspection, is captured and actioned.

Cargo Compliance Verification is randomly applied to consignments that would not typically be directed for inspection. Biosecurity officers look for biosecurity risk material — contamination with soil, animal or plant material — and check that importers have the required paperwork. They inspect the commodity, packing materials, and cleanliness of the internal and external surfaces of the container.

#### 7.1.2 Motivating Question

The motivating question for this study is how to compute appropriate performance indicators for the CCV exercise, including both point estimates (representing the best-supported estimate) and interval estimates (representing the statistical uncertainty associated with the point estimate).

As the aim of this sub-project is to provide advice for ongoing reporting, the results presented are illustrative in nature but present a range of possible approaches, as well as advice for future development of these reporting mechanisms.

---

\*Statistical Consulting Centre, The University of Melbourne

<sup>†</sup>Compliance Division, Department of Agriculture and Water Resources

<sup>1</sup>This chapter extends that presented in Clarke et al. (2015a).

<sup>2</sup><http://www.agriculture.gov.au/import/arrival/clearance-inspection/compliance-verification>

### 7.1.3 Methods

The data supplied for this sub-project were cargo compliance verification inspection results for the months of July and August, 2013. This included the line-level inspection results which indicated whether a line was inspected as part of CCV or otherwise, and the detailed (and potentially predictive) characteristics of each line including processing state, country, tariff, broker, importer and supplier. There was also line-level profile data, which identified those profiles that matched the line, and entry referral data, which indicated when cargo had a profile match at the entry level.

We adapted the work reported in Robinson et al. (2011) and Robinson et al. (2013) to better match the characteristics of the cargo pathway and the CCV program.

### 7.1.4 Results

The following chapter reports point and interval estimates for the nominated performance indicators.

### 7.1.5 Conclusions and Future Directions

The outcomes of this chapter have been superseded by CEBRA project 1501F, *Performance Indicators for Border Compliance* (Hoffmann et al., 2016).

## 7.2 Data preparation

The data supplied for this sub-project were cargo compliance verification inspection results for the months of July and August, 2013. This included the line-level inspection results which indicated whether a line was inspected as part of CCV or otherwise, and the detailed (and potentially predictive) characteristics of each line including processing state, country, tariff, broker, importer and supplier. There was also line-level profile data, which identified those profiles that matched the line, and entry referral data, which indicated when cargo had a profile match at the entry level. Here, *profiles* refer to filters in the Integrated Cargo System (ICS), which is managed by the Department of Immigration and Border Protection. The ICS profiles are the means by which consignments that are of interest to DAWR are identified and referred to AIMS. Profiles are constructed to capture consignments based on their economic tariff codes or the goods supplier, among other fields.

The only data cleaning issue was that there was inconsistency in the date coding, in particular, the use of a mixture of short and long dashes. All data preparation for this sub-project can be implemented in R should future data importing be performed (R Core Team, 2013).

## 7.3 Monthly CCV performance

The primary aim of this sub-project was summarising the monthly CCV results. As the aim was to be largely illustrative, and to supply the tools in R for the department to develop their own reporting system, only the results for July 2013 are reported here, though August 2013 data were also analysed to confirm repeatability.

The full data supplied contained all inspection results, so the initial data preparation involved extracting only those for the CCV project, that is, only those released after verification of documentation but still inspected. This analysis was also restricted to line entries only, and all food profiles were excluded.

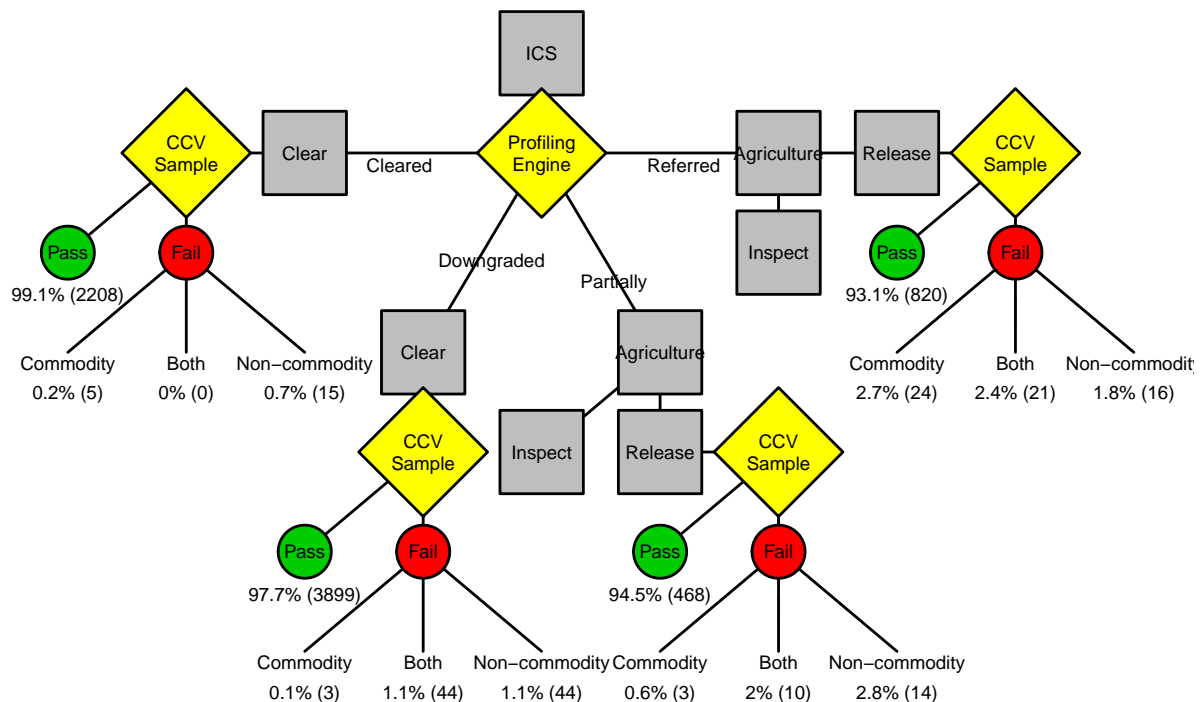
For each pathway or level, the error or failure rate based on CCV has been calculated, overall and for non-commodity, commodity and dual failure types separately.

Figure 7.1 provides a summary of the failures along each of the arms of the departmental profiling process, and splits the failures up by their nature, namely, commodity, non-commodity, or both.

Tables of results by a range of factors have been provided in Appendix E.3; Table 7.1 below is an example. The key results in these tables can also be presented as bar charts, such as Figure 7.2 for tariffs; bar charts for other characteristics have been provided in an appendix. All R scripts to produce these results have also been provided.

## 7.4 Monthly system-wide measures

A desirable feature of any performance summary is the ability to calculate system-wide measures like PIC (Post-Intervention Compliance) and BIC (Before Intervention Compliance) (Robinson et al., 2011,



**Figure 7.1:** The cargo compliance verification (CCV) summary for July 2013, with more detailed failure outlines.

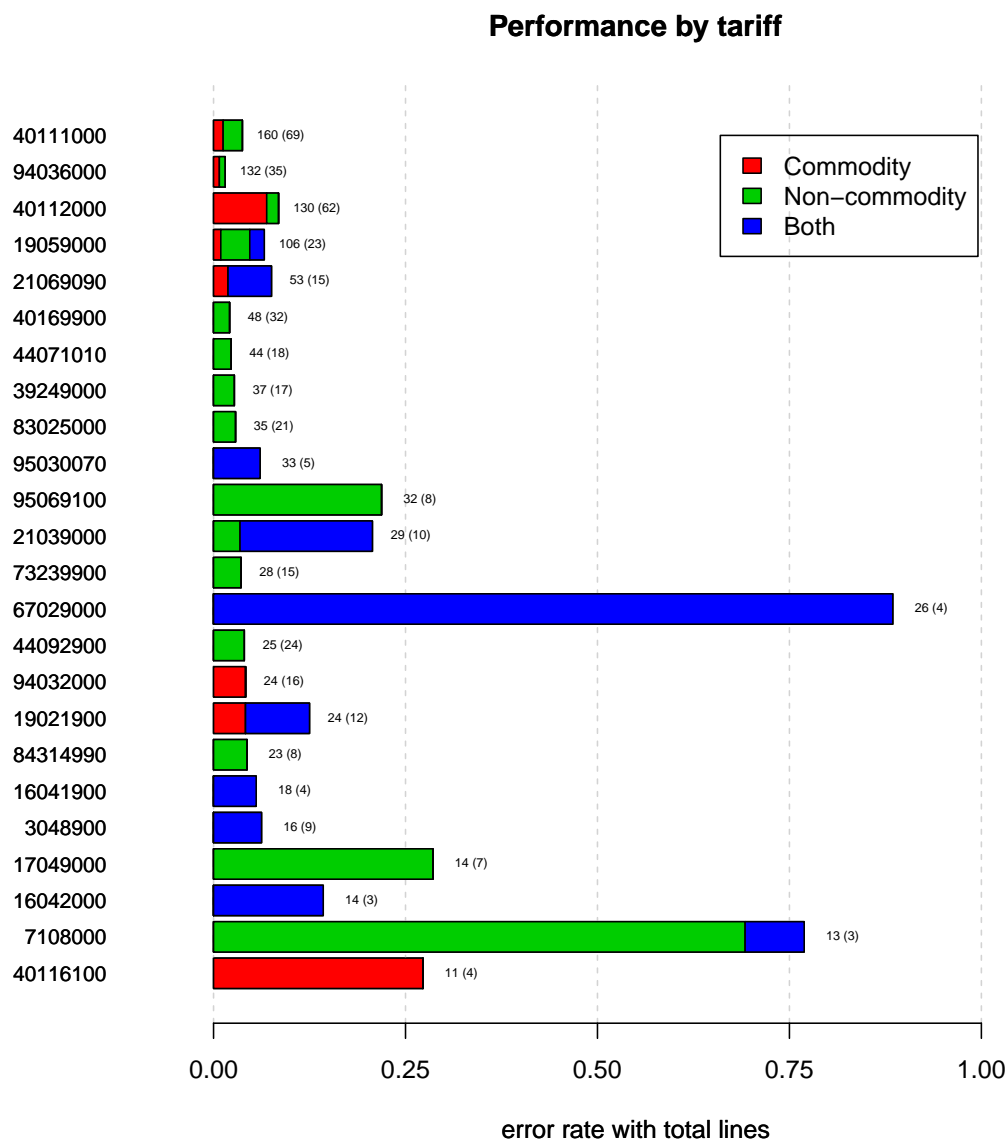
**Table 7.1:** CCV failure rates by country.

	Overall	Commodity	Non-Com	Both	Lines	Entries
IRAN	0.833	0.833			6	2
INDIA	0.127	0.054	0.074		204	37
MALAYSIA	0.097	0.014	0.083		72	28
AUSTRIA	0.091		0.091		22	8
JAPAN	0.056			0.056	719	31
NEW ZEALAND	0.052		0.052		248	55
CHINA	0.051	0.009	0.023	0.018	1859	294
CHILE	0.023		0.023		43	15
FRANCE	0.015		0.015		68	19
TAIWAN	0.012	0.004	0.008		256	40
INDONESIA	0.011		0.011		186	43
UNITED STATES	0.003		0.003		1218	59
THAILAND	0.003			0.003	307	46

2013). However, in order to do this, the total numbers within each pathway are required so that their contribution can be appropriately weighted. The total number of entries for the month of July were provided in order to estimate these measures and provide some indication of their precision in the form of a confidence interval.

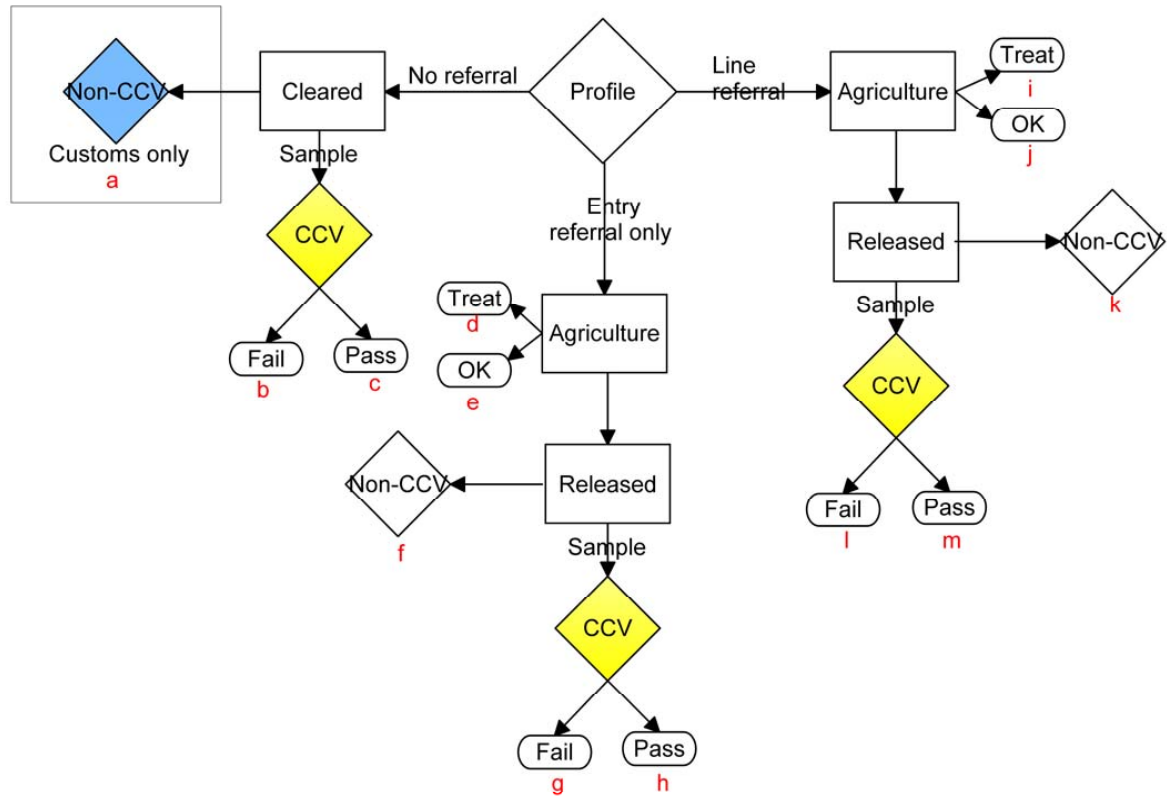
Figure 7.3 provides a summary of the pathways, indicating the data required from Customs. Letters have been assigned to each pathway node to represent the count of the number of consignments following the path to that node.

Table 7.2 gives the definitions, according to the data conventions used in the files provided, for the data in each node. As well as this, it gives the numbers of each (both entries and lines) for the month of July, node a excepted, for which only the entry level number was available. The R scripts used to calculate these values are contained in an appendix.



Results for the largest 24 levels with at least one failure out of a total of 100. There were a total of 1055 levels with no failures.

**Figure 7.2:** A bar chart demonstrating the failure rates for a range of tariff codes for July 2013 with lines (entries).



**Figure 7.3:** System-wide CCV flowchart with letters assigned to each node.

**Table 7.2:** Definition of nodes including counts for July 2013. A dash indicates that the datum for this column is not relevant to the calculation, and ‘NA’ is an actual level of the variable considered.

Nodes	Line Referral	Entry Referral	RoD	Inspected	Inspection OK?	Count Entries	Count Lines
a	N	N	Y	N	NA	84726	
b	N	N	Y	Y	N	27	172
c	N	N	Y	Y	Y	311	6818
d	N	Y	N	Y	N	32	235
e	N	Y	N	Y	Y	144	1346
f	N	Y	Y	N	NA	1052	6456
g	N	Y	Y	Y	N	3	4
h	N	Y	Y	Y	Y	35	365
i	Y	-	N	Y	N	591	2452
j	Y	-	N	Y	Y	3652	12108
k	Y	-	Y	N	NA	14686	53629
l	Y	-	Y	Y	N	33	168
m	Y	-	Y	Y	Y	543	1535



The final calculations for PIC and BIC as a percentage are given by the following formulae (and calculated for entries in July 2013):

$$\begin{aligned}
BIC &= 100 - 100 \left( \frac{a+b+c}{N} \times \frac{b}{b+c} + \frac{f+g+h}{N} \times \frac{g}{g+h} + \frac{d}{N} \right. \\
&\quad \left. + \frac{k+l+m}{N} \times \frac{l}{l+m} + \frac{i}{N} \right) = 92.1\%, \\
PIC &= 100 - 100 \left( \frac{a}{N} \times \frac{b}{b+c} + \frac{f}{N} \times \frac{g}{g+h} + \frac{k}{N} \times \frac{l}{l+m} \right) = 92.7\%,
\end{aligned}$$

where  $N = a + b + \dots + m$ .

## 7.5 Precision

When a statistic for a population is estimated using a sample of that population, there will always be a degree of uncertainty around how accurately that statistic reflects the true value. This is due to the natural variation in the sample population that arises from random selection.

When reporting a statistic calculated from a sample, uncertainty (or precision) is usually expressed as a confidence interval around the point estimate. Confidence intervals have an associated confidence level, which is usually 95%. This means that (other things being equal), if we were to repeat the sampling process twenty times over, then in nineteen of those twenty iterations (i.e. 95%) we will get results that enclose the true value.

For example, if 100 samples detect 5 units with leakage, then the point estimate of the leakage rate is 5%. However, the point estimate is reported with an associated 95% confidence interval of 2%–11%, which is an outcome of a process that captures the true value 95% of the time. For a total pathway population of 100 000, this would mean that the total number of leaked units is estimated at 5000, but that we are 95% confident that it is between 2000 and 11 000.

Confidence intervals only account for statistical variation arising from the random selection of samples; they do not account for non-statistical errors such as sample selection bias, imperfect inspection effectiveness, data entry errors, etc.

We do not advocate that confidence intervals be provided to managers routinely, but rather that they be used by analysts to advise managers as to the best way to achieve the trade-off between timeliness and detail (fine-grained, time-sensitive indicators) and stability and reliability (coarse, smoothed indicators). For example, the confidence interval might indicate that month-specific estimation of performance in smaller regions is too focused, that the sample sizes are too small, and that either (i) more samples should be taken, and/or (ii) the statistics should be computed based on the last quarter's values instead of the last month's values, and/or (iii) some of the values should be averaged across small regions.

**Recommendation 13.** CCV performance indicator interval estimates should be used by analysts to guide the temporal and hierarchical level of reporting of CCV performance indicator point estimates and to assist in the interpretation of the point estimates by pathway managers. The quality of the point estimates can be translated from the interval estimates to the pathway managers by the analysts (e.g., “the point estimate is poor”).

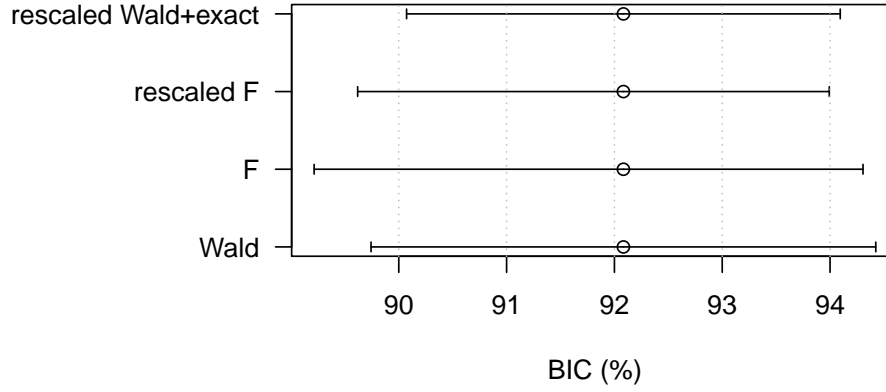
There are a variety of different methods that can be employed to produce confidence intervals for weighted sums of proportions; in this instance we have implemented the approaches detailed in Waller et al. (1994) which allow for cases where the proportions are near the extremes of 0 or 1. The confidence intervals for BIC and PIC are given in Figures 7.4 and 7.5, respectively, as well as Table 7.3. These results are fairly consistent; according to Waller et al. (1994), the rescaled F has been found to have the best coverage performance on simulated data.

Even without Customs system-level data, we could consider other summary measures. For example, we could compare the performance of the profiled lines or entries with those that are not referred, to indicate the usefulness of the profiling. For example, the odds ratio (OR) of contamination over non-contamination for the lines referred compared to those not referred (see Figure 7.3 for definitions of variables) is given by:

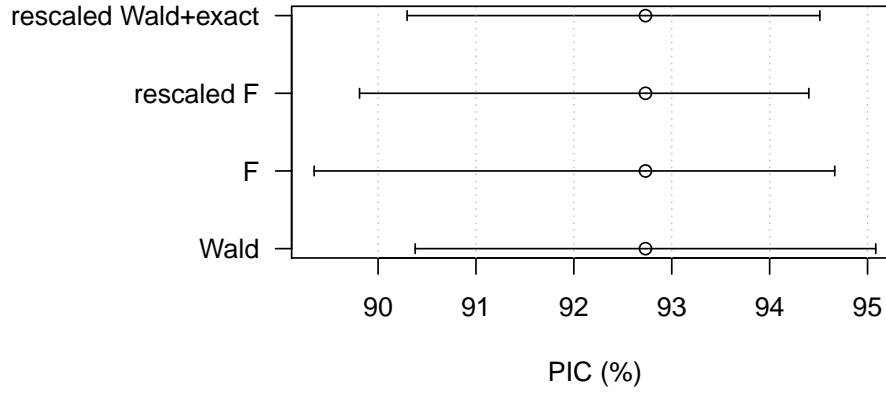
$$OR = \frac{l * c}{m * b}$$

**Table 7.3:** 95% confidence intervals for BIC and PIC, for entries in July 2013.

Method	BIC 95% CI	PIC 95% CI
Wald	(89.7%, 94.4%)	(90.4%, 95.1%)
F	(89.2%, 94.3%)	(89.3%, 94.7%)
rescaled F	(89.6%, 94.0%)	(89.8%, 94.4%)
rescaled Wald + exact	(90.1%, 94.1%)	(90.3%, 94.5%)



**Figure 7.4:** Confidence intervals for BIC under CCV, for entries in July.



**Figure 7.5:** Confidence intervals for PIC under CCV, for entries in July.

For the month of July, this odds ratio is 4.34 (95% CI: 3.48, 5.40). Hence we estimate that the odds of a failure of those referred using line profiles and released is 4.34 times higher than those not referred at all. We are 95% confident that the true odds ratio is between 3.48 and 5.40. This contrasts with the results for the month of August where the odds ratio was 0.81 (95% CI: 0.67, 0.99), indicating that the odds of a failure of those referred using line profiles was lower than for those not referred. This result shows that the variability from month to month can be considerable. The variability may arise from seasonal fluctuations in the types of consignments that are presented at the border, or from a change in the profiles.

## 7.6 Conclusion and Recommendations

As discussed, calculating system-wide measures requires Customs data of all cargo, not just those subject to profiling or CCV.

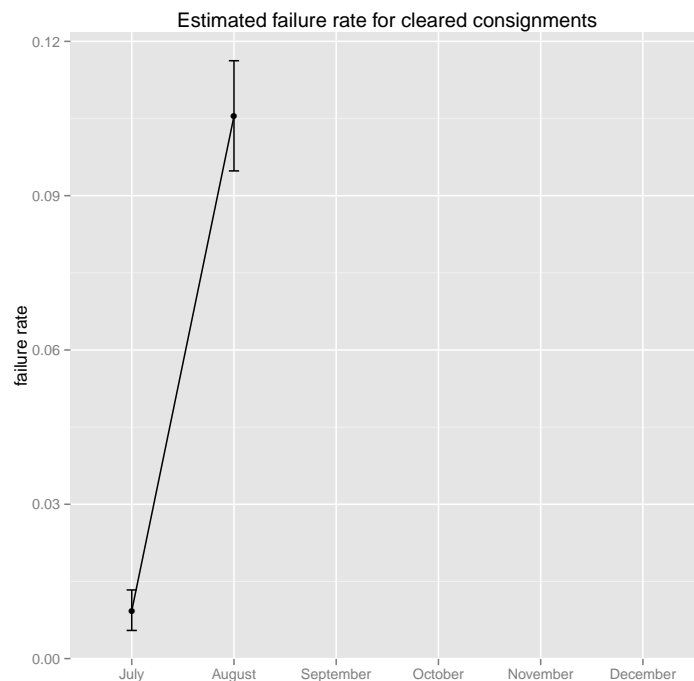
**Recommendation 14.** Performance indicators for CCV require information about the full range of goods imported to Australia. The department should obtain ICS data to enable computing system-wide measures of CCV performance indicators.

Discussions with the department established that it was also desirable to have summaries at the policy area level, which depend on the tariff codes of the individual lines. A code table relating policy areas and tariffs would be required to subset the results in this way, but the approach and potential outputs would be similar. An additional complication is the differential sampling rates of tariffs which would need to be considered in any policy level summaries.

**Recommendation 15.** CCV performance measures can be further distinguished by the policy area to which they are relevant by linking the profiles and tariffs to policy areas. The department should develop a code table that connects policy areas and tariffs to enable more fine-grained reporting for CCV performance measures.

Although here we consider reporting at the monthly level, there will of course be interest in, say, yearly or quarterly summaries, or analysis of changes over time. Yearly or quarterly summaries can follow similar approaches to those proposed here with combined data sets across broader time scales.

For analysis of changes over time, there are a range of approaches depending on the specific policy goal, but reporting approaches which provide a visual display of the results against time are going to be particularly worthwhile. Figure 7.6 is a possible way to present the results for a given pathway over time.



**Figure 7.6:** An example of a graphic designed to show changes over time, with estimate and 95% confidence interval based on CCV results.

**Recommendation 16.** This project computed CCV performance measures for just two months, as a demonstration of the underlying principles. The department should compute CCV performance measures for more time periods and consider reporting at a coarser scale, e.g. quarterly, or averaging across more than one month when reporting monthly.

## Chapter 8

# Interception Patterns of Hitch-hiker Pests

MATTHEW CHISHOLM\*

### 8.1 Summary

#### 8.1.1 Background

Hitchhiker pests are those that use a pathway as a means of transport but are not directly associated with the goods on the pathway, that is, the nature of the goods is not a factor. Examples of hitchhiker pests include: giant African snails and red imported fire ants arriving on shipping containers.

#### 8.1.2 Motivating Question

This case study was intended to test the hypothesis that the arrival of specific hitchhiker pests on imported cargo can be anticipated based on pest biology and known offshore distribution. If successful, the case study would have determined the viability of targeting imported cargo consignments more likely to contain specific hitchhiker pests, and may deliver an improved profiling of cargo based on non-conventional indicators of biosecurity risk.

#### 8.1.3 Methods

We tried to obtain suitable interception data from the department's AIMS and Incidents data holdings, but were unable to source sufficient data for satisfactory statistical analysis.

#### 8.1.4 Results

The study was withdrawn owing to a lack of suitable inspection data. However, considerable background work had been performed leading up to the withdrawal of the study, and this background work is reported here. Some of the material remains in relatively informal note form.

#### 8.1.5 Conclusions and Future Directions

The study was withdrawn owing to a lack of suitable inspection data.

### 8.2 Datasets

The Incidents database captures information on intercepted pests and diseases. Records from cargo border interceptions should contain the associated AIMS entry number and line number. The Bottle ID may also provide a way to link back to original records in AIMS, MAPS etc. There may be a degree of under-reporting of interceptions, i.e., an Incidents entry may not always be raised when a pest is found.

---

\*CEBRA, The University of Melbourne

Incidents raised from post-quarantine detections may include an AIMS reference, but only if it can be tied to a consignment, and even then it may not be recorded.

**Recommendation 17.** The hitch-hiker interception pattern analysis was significantly impeded by the lack of suitable interception data. Specific problems were inadequate linkages between the department's databases, and too few recorded instances of interception records for analysis. The department should develop more robust data capture and curation systems for gathering interception and operational data.

## 8.3 Organisms of interest

The original vision of the case study was to use two species, namely red imported fire ant (*Solenopsis invicta*), and guava rust (*Puccinia psidii*), as example hitchhiker pests, and target the inspection of consignments exported from areas where they are present, at times when they are reproducing or are more likely to be contaminating goods for export.

Other pests suggested that may be considered for a study of hitch-hikers include

- Brown marmorated stink bug *Halyomorpha halys*
- Emerald ash borer *Agrilus planipennis*
- Giant african snail *Achatina fulica*
- Asian gypsy moth *Lymantria dispar*
- Tramp ants more broadly
- Land snails.

DAWR entomologists should be able to suggest likely seasonal and locational patterns associated with hitch-hiker pests.

## 8.4 Potential organisations of interest

Preliminary research of literature and internet pages identified several organisations that have an interest in monitoring and/or controlling hitch-hiker pests. This research was performed in mid-2014, so may no longer be current.

### 8.4.1 United Nations FAO

The United Nations Food and Agriculture Organisation (FAO) website includes the following pages of possible interest.

- Comprehensive listings of government structures in Asia Pacific countries related to plant quarantine etc. (<http://www.fao.org/docrep/010/ag123e/AG123E00.htm#Contents>). The currency of this information has not been verified.
- A page for the IUFRO symposium on forest diseases and insects (<http://www.fao.org/docrep/24847e/24847e00.htm#Contents>) has a section on preventive measures, pathway characterisation etc., but this symposium was still in the early days of containerisation.
- A page for the proceedings of a workshop in 2003 on the 'Identification of risks and management of invasive alien species using the IPPC framework' (<http://www.fao.org/docrep/008/y5968e/y5968e00.htm#Contents>).

## 8.4.2 IPPC

The International Plant Protection Convention (IPPC, <https://www.ippc.int/>) is an international agreement on plant health to which 181 signatories currently adhere. It aims to protect cultivated and wild plants by preventing the introduction and spread of pests. The Secretariat of the IPPC is provided by the FAO.

The IPPC has been developing a draft ISPM ‘Minimizing pest movement by sea containers’, the most recent available draft of which is at [https://www.ippc.int/sites/default/files/documents/20131011/2008-001\\_draft\\_ispm\\_seacontainers\\_en\\_2013-06-26\\_outofocs\\_2013101109%3A35--105KB.doc](https://www.ippc.int/sites/default/files/documents/20131011/2008-001_draft_ispm_seacontainers_en_2013-06-26_outofocs_2013101109%3A35--105KB.doc). This document is being developed by the Expert Working Group on Sea Containers (EWGSC). Their webpage (<https://www.ippc.int/core-activities/standards-setting/expert-drafting-groups/expert-working-groups/sea-containers>) provides links to all relevant documents and webpages, such as the history of document’s development, meetings where it was discussed, minutes of those meetings etc.

The lead steward of the document is John Hedley, Principal Adviser—International Organizations, New Zealand Ministry for Primary Industries. Mr Hedley is also a member of the Standards Committee (SC), which is a subsidiary body of the Commission on Phytosanitary Measures (CPM). The current SC membership is available at [https://www.ippc.int/sites/default/files/documents/20140417/membership\\_sc\\_contactinfo\\_2014-04-17\\_201404171612--181.79KB.pdf](https://www.ippc.int/sites/default/files/documents/20140417/membership_sc_contactinfo_2014-04-17_201404171612--181.79KB.pdf).

The draft ISPM has been developed over several years, and was provided to all IPPC member states and stakeholder organisations for comment in 2013. The compiled comments from this exposure are available at <https://www.ippc.int/publications/2013-compiled-comments-draft-ispm-minimizing-pest-movement-sea-containers-2008-001>. Many of the comments indicate serious concerns with the document, including its unclear scope, inconsistent terminology, unclear areas of responsibility, over-reliance on manual inspection, the expectation that 100% of containers be inspected on all six sides, the expectation that records be kept for all inspections, and hence its impact on trade.

One of the main concerns among the compiled comments is the lack of a risk-based justification. This is now the subject of an international survey of containers to determine the level of contamination. According to the comments, the idea to perform a survey was endorsed at CPM-8 in May 2013. The minutes of the SC meeting, available at [https://www.ippc.int/sites/default/files/documents/20130618/report\\_sc\\_2013\\_may\\_xxii\\_2013-06-17\\_2013061810%3A14--1.81MB.pdf](https://www.ippc.int/sites/default/files/documents/20130618/report_sc_2013_may_xxii_2013-06-17_2013061810%3A14--1.81MB.pdf), indicate that

- the survey is being coordinated by a small group, of which John Hedley is a member
- NPPOs would be asked to complete the survey for as many sea containers as possible and, if possible, for all six sides of the containers
- data would not be gathered on the origin of the containers, or on the volume of the various contaminations found, due to the availability/complexity of these data
- NPPOs may also decide to survey the inside of containers, in addition to the outside
- the data collected would be used as a baseline for measuring the impact of future protocols
- the survey design would be finalised by mid July 2013, the survey would be performed Sep–Dec 2013, and that the results would be analysed by Feb 2014.

The comments provided by the World Shipping Council (WSC) are particularly thorough and robust. The motivation of the CEBRA hitch-hikers sub-project is reinforced by their comment on page 32:

*documentation of any substantial risk of plant pest from particular geographic locations should be systematically obtained together with the identification of those risks and other relevant pest management data, e.g., time of year of prevalent risk of infestation, so that authorities and industry could understand the risk at issue and could develop appropriate, specific remedial responses.*

Comments, particularly by the WSC and the World Customs Organisation, also provide valuable insight into the details of global container logistics.

The CPM met again in April 2014. According to the EWGSC’s webpage, at this meeting the CPM

- “recognised and appreciated the joint initiative by the International Maritime Organization (IMO), the International Labour Organization (ILO) and United Nations Economic Commission for Europe

(UNECE) of revising the Code of Practice for Packing of Cargo Transport Units (CTU Code). With the support from the IPPC Expert Working Group on Sea Containers, those organizations have incorporated into the revised CTU Code several elements of phytosanitary relevance, e.g. information on pests and other contamination which may be associated with CTUs, as well as very useful practical guidelines for cleanliness, cleaning, packing and handling”

- “welcomed the recent adoption of the CTU Code by UNECE and looked forward to the adoption also by IMO and ILO of the revised CTU Code later this year”
- “emphasised that the careful implementation of the revised CTU Code by all operators responsible for and involved in the packing and handling of sea containers is crucial for preventing the spread of pests and invasive alien species.”

The enhanced Code of Practice is at [http://www.unece.org/fileadmin/DAM/trans/doc/2014/itc/id\\_07\\_CTU\\_Code\\_January\\_2014.pdf](http://www.unece.org/fileadmin/DAM/trans/doc/2014/itc/id_07_CTU_Code_January_2014.pdf). As this code is implemented, there may be a reduction in the observed contamination rate of containers.

### 8.4.3 Regional Plant Protection Organisations

A Regional Plant Protection Organization (RPPO) is an inter-governmental organization functioning as a coordinating body for National Plant Protection Organizations (NPPO) on a regional level. Not all contracting parties to the IPPC are members of RPPOs, nor are all members of RPPOs contracting parties to the IPPC. Moreover, certain contracting parties to the IPPC belong to more than one RPPO. There are currently ten RPPOs. (Source: <https://www.ippc.int/partners/regional-plant-protection-organizations>)

**APPPC** The Asia and Pacific Plant Protection Commission (APPPC) does not include any Pacific Island Countries and Territories (PICTs), but does include the Indian subcontinent. PICTs are instead represented in the Pacific Plant Protection Organization (see below). The APPPC’s website (<http://www.apppc.org/>), contains the following documents of interest.

- The *Report of the 14th APPPC Regional Consultation on Draft ISPMs* ([http://www.apppc.org/sites/apppc.org/files/1384223695\\_Report\\_of\\_the\\_14th\\_APPPC\\_Regional.doc](http://www.apppc.org/sites/apppc.org/files/1384223695_Report_of_the_14th_APPPC_Regional.doc)), which includes two pages of notes from the discussions on the draft ISPM ‘Minimizing pest movement by sea containers’. From these notes it appears that the main member countries concerned with preventing hitch-hiker imports are Australia, NZ, USA, Canada and Chile. Other countries, particularly China, Japan and Korea, appear more focused on preventative measures at the export end. This meeting was attended by John Hedley, but Australia did not send a representative. The next meeting will be in Pusan (Korea) on 15–19 September 2014 (<http://www.apppc.org/content/15th-apppc-regional-workshopippc-workshop-review-draft-ispms-pusan-republic-korea-15-19>).
- The Draft Regional Standards for Phytosanitary Measures ([http://www.apppc.org/sites/apppc.org/files/1219050545298\\_APPPC\\_Draft\\_RSPM\\_containers\\_0.doc](http://www.apppc.org/sites/apppc.org/files/1219050545298_APPPC_Draft_RSPM_containers_0.doc)) appears to be a fore-runner of the current draft ISPM (see above). The document includes a list of official terminology, and reference to a 2006 MAF Biosecurity New Zealand document titled ‘Monitoring Research and Pathway review: Sea Containers’.

**EPPO** The European and Mediterranean Plant Protection Organization (EPPO) includes all states of Europe except Montenegro & Iceland, plus all Mediterranean littoral states except Libya, Egypt, Syria & Lebanon, plus Azerbaijan, Kazakhstan, Uzbekistan & Kyrgyzstan. The EPPO’s website (<http://www.eppo.int/>) includes the following pages of potential interest.

- Brief notes on the *55th Meeting of the Panel on CPM Affairs—Joint EPPO/NAPPO Meeting* ([http://www.eppo.int/MEETINGS/2014\\_meetings/cpm\\_montreal.htm](http://www.eppo.int/MEETINGS/2014_meetings/cpm_montreal.htm)) mention that the draft ISPM on sea containers was discussed, including “the addition of more data through statistical analysis of existing data, [and the] feasibility of an additional survey”. This page includes an attendee list. There are several other pages minuting earlier meetings that covered the draft ISPM.
- There are pages for the *EPPO A1 and A2 List of pests recommended for regulation as quarantine pests* (<http://www.eppo.int/QUARANTINE/listA1.htm>, and <http://www.eppo.int/QUARANTINE/listA2.htm>). These pages contain comprehensive lists of pests, along with papers on their distribution, biology etc.

**NAPPO** North American Plant Protection Organization (NAPPO) has a website (<http://www.nappo.org/>) which was inaccessible at each attempt. There is a related website for the NAPPO's Phytosanitary Alert System (<http://www.pestalert.org/main.cfm>), which 'provides up-to-date information on pest situations of significance to North America'. This website has many official pest alerts by USA and Canada, plus some by Mexico, as well as a smaller number of unofficial reports about organisms around the world. It also provides a page of links (<http://www.pestalert.org/resources.cfm>) to many other websites about pests.

**COSAVE** Comité de Sanidad Vegetal del Cono Sur (COSAVE) represents states of the South American cone: Argentina, Bolivia, Brazil, Chile, Paraguay, Peru and Uruguay. Their website (<http://www.cosave.org/pagina/bienvenidos-al-comite-de-sanidad-vegetal-cosave>, Spanish only) is comprehensive, and includes a current list of regulated pests in the region (<http://www.cosave.org/pagina/listado-de-las-principales-plagas-reglamentadas-para-la-region-del-cosave>). The list mentions the member countries where the pest is present, but doesn't give detailed information on distributions.

**Other RPPOs** The websites of the following RPPOs either did not yield any useful leads, or were not investigated because they consist primarily or entirely of contiguous states, for which sea container hitch-hikers would be largely irrelevant.

- Comunidad Andina (CA)—Representing four states of the Andes
- Caribbean Plant Protection Commission (CPPC)—Most island states and some littoral states of the Caribbean
- Inter-African Phytosanitary Council (IAPSC)—All African Union members except Morocco
- Near East Plant Protection Organization (NEPPO)—Primarily states in the Middle East & Maghreb
- Organismo Internacional Regional de Sanidad Agropecuaria (OIRSA)—Primarily Central American states
- Pacific Plant Protection Organization (PPPO)—PICTs plus Australia, New Zealand, USA and France (for French Polynesia).

#### 8.4.4 Secretariat of the Pacific Community

The Secretariat of the Pacific Community (SPC, <http://www.spc.int/>) is the Pacific Island region's principal technical and scientific organisation. It delivers technical, scientific, research, policy and training support to Pacific Island countries and territories in fisheries, agriculture, forestry, water resources, geo-science, transport, energy, disaster risk management, public health, statistics, education, human rights, gender, youth and culture.

It has 26 member countries and territories including its founding members, Australia, France, New Zealand and the USA. Its headquarters are in Noumea, New Caledonia, with other offices in Fiji, Federated States of Micronesia and Solomon Islands.

The SPC's Biosecurity and Trade section is part of its Land Resources Division, which is based in Suva, Fiji, and has approximately 74 staff. This organisation appears to subsume the PPPO in some ways—its website is much more informative than that of the PPPO, and it 'hosts the PPPO Secretariat, as it is the regional organisation that hosts all member countries and is involved in providing assistance in plant protection and quarantine to member countries'. Information on the PPPO, and on the PICTs' attendance at the IPPC's 9th CPM, actually appear on the SPC's website.

The SPC's website contains useful information, and the organisation may prove to be a valuable collaborator on the project. One of their related websites is the Pacific Islands Pest List Database (<http://www.spc.int/pld/>), which provides comprehensive information on pests of interest in the region, including current distributions and known hosts in PICTs.

#### 8.4.5 Pacific Islands Development Program

The Pacific Islands Development Program (PIDP) conducts a broad range of activities to enhance the quality of life in the Pacific islands. It is part of the East-West Center, based in Honolulu. It published weekly digests of news of interest to its members, one of which mentions the Micronesia Biosecurity Plan (MBP, <http://pidp.eastwestcenter.org/pireport/2012/January/01-20-05.htm>).



The MBP is described as “a unique partnership of U.S. agencies . . . with a variety of local and regional . . . experts”, as “a proactive effort that will analyze risks of various pathways, organism [sic.] and vectors to Guam to other areas of Micronesia associated with the pending [U.S.] military relocations to Guam and the [Northern Marianas]”.

According to a page on the Governor of Hawaii’s website (<http://governor.hawaii.gov/blog/aberrant-administration-requests-federal-recognition-of-hawaiis-unique-biosecurity-needs/>), “The U.S. Department of Defense is developing the Micronesian Biosecurity Plan (MBP) in preparation for potential military relocations in the Pacific region, recognizing the potential for the accidental transport of invasive species. The MBP is both a risk assessment of potential invasive species pathways in the Pacific and a set of recommendations for enhancing Pacific biosecurity.”

#### 8.4.6 CIRAD

The *Centre de coopération internationale en recherche agronomique pour le développement* (CIRAD, <http://www.cirad.fr/en/>) is a French research centre working with developing countries to tackle international agricultural and development issues. It does research for francophone jurisdictions in the western Indian Ocean and the Caribbean, specifically Comoros, Madagascar, Mauritius, Seychelles, Réunion, Mayotte, French Guiana, Guadeloupe and Martinique.

#### 8.4.7 Agrisles

Agrisles (<http://www.agrisles.eu/en/accueil.html>) aims ‘to create a sustainable system of closer relationship among public authorities in order to better coordinate their search for solutions to similar problems in their agriculture, and thus improve, altogether, the efficiency of regional policies. But it is also to help economic, technical and scientific actors of the different islands to meet each other in this framework.’ It represents the Azores, Balearic Islands, Corsica, Sardinia, the Greek archipelagos, Malta and Cyprus. No mention was found of container sanitation. Member countries’ quarantine laws would all be subsidiary to the European Union.

#### 8.4.8 Organisation for Economic Cooperation and Development (OECD)

In 2013 Andrew Robinson and Andrew Liebhold submitted a proposal to the OECD’s Cooperative Research Programme for funding to hold a week-long *Workshop on Hitchhiking Invasion Pathways: Toward Exclusion of Invading Species Damaging to Agriculture and Forestry*. While the funding bid was unsuccessful, the groundwork could be rolled into this project, and the proposed attendees may be willing to collaborate.

#### 8.4.9 European Union

The European Union (EU) has the following webpages of interest.

- Monthly and annual reports of interceptions of harmful organisms in imported plants and other objects ([http://ec.europa.eu/food/plant/plant\\_health\\_biosafety/europhyt/interceptions\\_en.htm](http://ec.europa.eu/food/plant/plant_health_biosafety/europhyt/interceptions_en.htm))
- Council directive on protective measures against the introduction into the Community of organisms harmful to plants or plant products and against their spread within the Community (<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2000:169:0001:0112:EN:PDF>)
- Food import conditions ([http://ec.europa.eu/food/international/trade/docs/decision\\_318\\_summury\\_en.pdf](http://ec.europa.eu/food/international/trade/docs/decision_318_summury_en.pdf)), but with no mention of containers or conveyances.

#### 8.4.10 National organisations

**Papua New Guinea** National Agriculture Quarantine and Inspection Authority (<http://www.naqia.gov.pg/>).

**Solomon Islands** Solomon Islands Customs and Excise Division (SICED, <http://www.mof.gov.sb/AboutUs/CustomsDivision.aspx>) and Solomon Islands Agricultural Quarantine Service (SIAQS, no website found). The Regional Assistance Mission to Solomon Islands (RAMSI, <http://www.ramsi.org/>) is a mission created in 2003 in response to a request for international aid by the SI Governor-General after significant regional unrest. Main involvement is by the Australian Federal Police and the Australian Defence Force. The ADF's involvement, under Operation Anode, ended in 2013. RAMSI deliverables don't appear to include anything related to quarantine, but they may have some data relating to hitchhikers found on containerised materiel and aid.

**New Caledonia** *L'Institut agronomique néo-Calédonien* (<http://www.iac.nc/>) has research priorities for the protection of natural resources and biodiversity in New Caledonia.

**Fiji** Biosecurity Authority of Fiji (<http://www.biosecurityfiji.com/about-us/our-organisation.html>) aims 'to be respected as the most effective and efficient Biosecurity Authority in the Pacific region'. This agency also hosts the forum *Heads of Quarantine in the Pacific* (see press release at <http://www.biosecurityfiji.com/media-centre/press-releases/143-bilateral-talks-amongst-heads-of-quarantine-boosts-trade-in-the-region.html>).

**Samoa** Samoa Quarantine Service (<http://www.samoaquarantine.gov.ws/>).

**Tonga** Quarantine & Exports Of Tonga (<http://www.quarantine.gov.to/index.php>).

**New Zealand** Ministry for Primary Industries (<http://mpi.govt.nz/>) has a specific page about importing containers and cargo (<http://www.biosecurity.govt.nz/regs/cont-carg>).

**Chile** *Servicio Agrícola y Ganadero* (<http://www.sag.cl/>) inspects all conveyances on arrival, but nothing is mentioned about containers. They have an active AGM program (<http://www.sag.gob.cl/ambitos-de-accion/lymantria-dispar-o-polilla-gitana>).

**Argentina** *Servicio Nacional de Sanidad y Calidad Agroalimentaria* (<http://www.senasa.gov.ar/index.php>), including the *Dirección Nacional de Protección Vegetal* office (<http://www.senasa.gov.ar/contenido.php?to=n&in=614&io=2417>).

**Brazil** In the Ministry of Agriculture, the Plant Health Department (*Departamento de Sanidade Vegetal*, <http://www.agricultura.gov.br/vegetal/sanidade-vegetal>) manages and runs phytosanitary activities for imports, including prevention of pest incursions.

**USA** Animal and Plant Health Inspection Service (<http://www.aphis.usda.gov/wps/portal/aphis/home/>).

**Canada** Canadian Food Inspection Agency (<http://www.inspection.gc.ca/eng/1297964599443/1297965645317>) performs risk analysis and sets policy, while the Canadian Border Services Agency (<http://cbsa-asfc.gc.ca/menu-eng.html>) enforces the requirements at all border points. The CFIA's directive D-95-26: *Phytosanitary requirements for soil and related matter, and for items contaminated with soil and related matter* (<http://www.inspection.gc.ca/plants/plant-protection/directives/imports/d-95-26/eng/1322520617862/1322525397975>) includes containers in its scope, but only in relation to soil contamination.

**Iceland** Icelandic Food and Veterinary Authority (Matvælastofnun, <http://www.mast.is/english/>).

**Ireland** Department of Agriculture, Food and the Marine, but quarantine policies (<https://www.agriculture.gov.ie/agri-foodindustry/tradeimportsexports/>) are subordinate to the EU.

**UK** Plant Health and Seeds Inspectorate (<http://www.fera.defra.gov.uk/plants/plantHealth/>), part of the Food and Environment Research Agency in the Department for Environment, Food and Rural Affairs. No references found to container sanitation requirements. Subordinate to EU.

**Malta** Plant Health Surveillance and Inspectorate Unit (<http://agric.gov.mt/surveillance-and-inspectorate?l=1>) performs physical inspections of imports. Subordinate to EU.

**Israel** Plant Protection and Inspection Services (<http://www.moag.gov.il/agri/English/Ministries+Units/Plant+Protection+and+Inspection+Services/default.htm>), part of the Ministry of Agriculture & Rural Development.

**Iran** Plant Protection Organisation (<http://www.ppo.ir/English/Pages/Default.aspx>).

**Mauritius** Within the Ministry of Agro Industry and Food Security, the Plant Quarantine section (<http://agriculture.gov.mu/English/Pages/Services/HealthandQualityCertification/Plant-Quarantine.aspx>).

**Sri Lanka** National Plant Quarantine Service (<http://www.agridept.gov.lk/index.php/en/institutes/338>) in the Department of Agriculture.

**Singapore** Agri-Food and Veterinary Authority (<http://www.ava.gov.sg/>). Based on their 2012 Annual Report, there appears to be minimal inspection of imports.

**Hong Kong** Agriculture Fisheries and Conservation Department (<http://www.afcd.gov.hk/english/quarantine/quarantine.html>).

**China** General Administration of Quality Supervision, Inspection and Quarantine (AQSIQ, <http://english.aqsic.gov.cn/>). Website was very slow to load, so was not investigated further.

**Korea** Animal and Plant Quarantine Agency (<http://www.qia.go.kr/english/html/indexqiaEngNoticeWebAction.do?clear=1>).

**Japan** Ministry of Agriculture, Forestry and Fisheries—Plant Protection Station (<http://www.pps.go.jp/english/index.html>).

**Taiwan** The Bureau of Animal & Plant Health Inspection & Quarantine (BAPHIQ, <http://www.baphiq.gov.tw/index/index.html>link). Inspects imports, but no mention of container sanitation. Not party to IPPC due to diplomatic status.

**Philippines** Plant Quarantine Service (<http://pqs.da.gov.ph/>) ‘aims to prevent the entry of foreign pests into the country, prevent spread of pests already existing in the country and comply with the international standards’. Information available on import inspection procedures (<http://pqs.da.gov.ph/index.php/procedures/inspection>) doesn’t mention containers. However, they sent a representative to the 14th APPPC regional meeting, which discussed the draft ISPM on sea containers, so may have unpublished requirements.

**Dead ends** The following government agency websites did not yield useful information:

- Vanuatu – <http://www.gov.vu/index.php/government/agriculture>
- Cyprus – [http://www.moa.gov.cy/moa/da/da.nsf/index\\_en/index\\_en?OpenDocument](http://www.moa.gov.cy/moa/da/da.nsf/index_en/index_en?OpenDocument)
- Nauru – <http://www.naurugov.nr/government/departments/department-of-justice-and-border-control/quarantine-section.aspx>
- Faroe Islands – [http://www.hfs.fo/portal/page?\\_pageid=33,42505,33\\_42563&\\_dad=portal&\\_schema=PORTAL](http://www.hfs.fo/portal/page?_pageid=33,42505,33_42563&_dad=portal&_schema=PORTAL)
- Greenland – <http://naalakkersuisut.gl/en/Naalakkersuisut/Departments/Fiskeri-Fangst-og-Landbrug>
- Maldives – <http://www.fishagri.gov.mv/index.php/en/>.

## 8.5 Sub-national organisations

**Tasmania** Department of Primary Industries, Parks, Water and Environment (DPIPWE, <http://dPIPWE.tas.gov.au/biosecurity-quarantine/quarantine-tasmania>). Nothing apparent on container sanitation.

**Antarctica** The Australian Antarctic Division (<http://www.antarctica.gov.au/>) is responsible for protecting, administering and researching the Australia Antarctic Territory and Australia's subantarctic islands (Heard, McDonald and Macquarie). Research was conducted in 2007–09 on pathways for the introduction of non-native species into Antarctica and the subantarctic.

**Hawaii** The Hawaii Invasive Species Council (HISC, <http://dlnr.hawaii.gov/hisc/>) is an inter-departmental collaboration of the Departments of Land & Natural Resources (DLNR), Agriculture (DOA), Health (DOH), Transportation (DOT), Business, Economic Development & Tourism (DBEDT), and the University of Hawaii (UH). The HISC was established in 2003 for the special purpose of providing policy level direction, coordination, and planning among state departments, federal agencies, and international and local initiatives for the control and eradication of harmful invasive species infestations throughout the State and for preventing the introduction of other invasive species that may be potentially harmful. The U.S. Plant Protection Act of 2000 prevents the state from regulating in foreign and interstate commerce pests that may threaten Hawaii but are not federally listed as threats to U.S. agriculture.

**Galapagos Islands** The Galapagos Inspection and Quarantine System (SICGAL, its acronym in Spanish) represents the primary barrier against future biological introductions. SICGAL was initiated in May 1999 and formally established in June 2000. A program of the Ecuadorian Service for Agricultural Health, SICGAL involves a high degree of inter-institutional coordination and cooperation, with the goal of preventing new species and organisms from being introduced into the Galapagos Islands. SICGAL involves inspections both on the continent and in Galapagos of flights, cargo ships, passengers, and cargo. Source: <http://www.galapagos.org/conservation/biosecurity/>. SICGAL's website ([http://sicgal.fundargalapagos.org/index.php?option=com\\_frontpage&Itemid=1](http://sicgal.fundargalapagos.org/index.php?option=com_frontpage&Itemid=1)) doesn't mention container sanitation among its import conditions.

## 8.6 Search method

Using Google and the University of Melbourne library, searches were conducted of websites and publications containing a number of key terms. In addition to the websites and documents that those searches identified, citations and references were also investigated, as well as other documents within the same internet domain.

To identify organisations of potential interest, search terms such as 'agriculture department', 'quarantine department' and 'biosecurity department' were combined with the names of countries, sub-national jurisdictions and known international agencies that might reasonably be expected to deal with significant volumes of sea container trade, and/or have an interest in preventing pest incursions.

Search terms include:

- sea container
- container contamination
- hitchhiker pest
- stowaway pest
- *fourmis auto-stoppeurs*
- red fire ant
- solenopsis invicta
- RIFA
- puccinia psidii
- myrtle rust
- asian gypsy moth
- AGM
- lymantria dispar
- ant worldwide spread
- alien species shipping vector

Most literature identified dealt with the distributions of pests, potential new habitats, and how those pests spread upon establishment. Few articles were identified that dealt with the vectors of travel. Searches for ‘shipping stowaways’ or variations thereof mostly identified literature about ballast water and/or biofouling, but very little about shipping containers.

## 8.7 Other projects

**Chevron Australia** The Gorgon project is developing the Gorgon and Jansz-Lo gas fields, including the construction of an LNG plant on Barrow Island. Strict quarantine protocols have been put in place for the island during the construction phase ([http://www.chevronaustralia.com/docs/default-source/default-document-library/gorgon\\_ch12\\_lr.pdf?sfvrsn=0](http://www.chevronaustralia.com/docs/default-source/default-document-library/gorgon_ch12_lr.pdf?sfvrsn=0)). Protocols include full inspection and cleaning of all sides of all containers prior to shipping from the mainland.

### CEBRA projects

- 1303A *Intelligence gathering and analysis* aims to provide real-time intelligence on emerging pests, diseases and pathogens. The main deliverable, the International Biosecurity Intelligence System (IBIS) web search tool, may provide current information on the distribution of pests that are prone to hitch-hike.
- 1302A *Evaluation of arrival pathways and species distribution models* is a scoping study of spatial analysis models for the purpose of providing ‘a sufficiently accurate assessment of species distribution, potential establishment sites or high-risk pathways’.
- 1305B *Plant-product pathways and the Continuous Sampling Plan* may present a model for future inspections of containers for hitch-hiker pests.

**Satellite Applications Catapult** Satellite Applications Catapult (<https://sa.catapult.org.uk/>) is a new UK government funded research company, beginning to work on a project that involves creating risk maps for *Lobesia botrana* in Chile. They are interested in creating a model for risk mapping of areas that have not yet been affected, and are looking for experienced collaborators and/or information. (pers. comm. A. Robinson, forwarded from Frank Koch email to IPRMW members 11 Jul 14).

# Literature Cited

- Argyriou, A., Maurer, A., and Pontil, M. (2008). An algorithm for transfer learning in a heterogeneous environment. In *Machine Learning and Knowledge Discovery in Databases*, pages 71–85. Springer.
- Clarke, S., Hollings, T., Liu, N., Hood, G., and Robinson, A. (2017). Biosecurity risk factors presented by international vessels: a statistical analysis. *Biological Invasions*, 19(10):2837–2850.
- Clarke, S., Hood, G., and Robinson, A. P. (2015a). Data Mining: Report on First Cohort of Case Studies. Technical Report 1301A2, Centre of Excellence for Biosecurity Risk Analysis.
- Clarke, S., Robinson, A. P., and Hood, G. (2014). Data Mining: Sub-Project Data Resources. Technical Report 1301A1, Centre of Excellence for Biosecurity Risk Analysis.
- Clarke, S., Stenekes, N., Kancans, R., Woodland, C., and Robinson, A. (2015b). Red letters and where they are going. In *IEEE International Symposium on Big Data Visual Analytics, Hobart, September 22–25 2015*.
- Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181.
- Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The Elements of Statistical Learning*. Springer, 2nd edition.
- Hoffmann, M., Robinson, A. P., and Holliday, J. (2016). Performance Indicators for Border Compliance. Technical Report 1501F1, Centre of Excellence for Biosecurity Risk Analysis.
- Hothorn, T. (2013). CRAN Task View: Machine Learning  
<http://cran.r-project.org/web/views/machinelearning.html>.
- Lim, J. J. (2012). *Transfer learning by borrowing examples for multiclass object detection*. PhD thesis, Massachusetts Institute of Technology.
- Miller, H., Clarke, S., Lane, S., Lonie, A., Lazaridis, D., Petrovski, S., and Jones, O. (2009). Predicting customer behaviour: The University of Melbourne’s KDD Cup report. *The 2009 Knowledge Discovery in Data Competition (KDD Cup 2009) Challenges in Machine Learning, Volume 3*, page 43.
- Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22:1345–1359.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. Y. (2007). Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM.
- Robinson, A., Cannon, R., and Mudford, R. (2011). AQIS Quarantine Operations Risk Return: Performance Indicators. Technical Report 1001I 1, Australian Centre of Excellence for Risk Analysis.
- Robinson, A., Chisholm, M., Mudford, R., and Maillardet, R. (2016). Ad-hoc solutions to estimating pathway contamination rates using imperfect and incomplete information. In Jarrad, F., Low-Choy, S., and Mengersen, K., editors, *Biosecurity Surveillance: Quantitative Approaches*, chapter 9, pages 167–180. CABI, Wallingford, Oxfordshire.

- Robinson, A., Mudford, R., Quan, K., Sorbello, P., and Chisholm, M. (2013). Adoption of meaningful performance indicators for quarantine inspection performance. Technical Report 1101D 1, Australian Centre of Excellence for Risk Analysis.
- Rosenstein, M. T., Marx, Z., Kaelbling, L. P., and Dietterich, T. G. (2005). To transfer or not to transfer. In *In NIPS'05 Workshop, Inductive Transfer: 10 Years Later*.
- Torrey, L. and Shavlik, J. (2009). Transfer learning. *Handbook of Research on Machine Learning Applications. IGI Global*, 3:17–35.
- Waller, J. L., Addy, C. L., Jackson, K. L., and Garrison, C. Z. (1994). Confidence intervals for weighted proportions. *Statistics in Medicine*, 13(10):1071–1082.

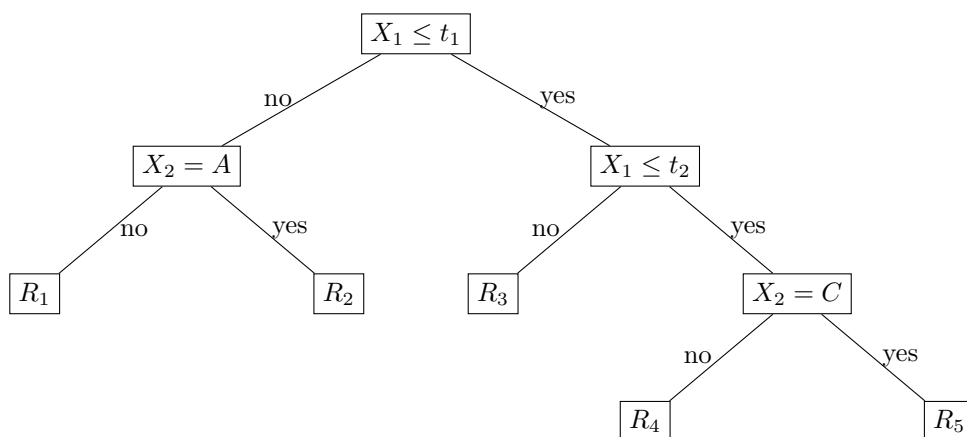
# Appendix A

## Random forests overview

Many of these sub-projects (and data mining approaches more generally) incorporate some kind of random forest modelling. This appendix provides a brief overview of such approaches.

In order to understand random forests—that is, collections of tree models—we need to first understand tree-based methods in general. Tree-based methods are conceptually simple but enable complex modelling. As with any model, there are candidate predictors which are used to attempt to predict an outcome. In the case of trees, the outcome we want to predict is usually either binary, such as the presence or absence of biosecurity risk material, or numerical, such as a count of the number of seizures in a particular region. For each observation (say, a mail item, a seizure or a spatial region), we observe this outcome and we also observe a set of characteristics of this observation, known as predictors. These predictors can be both categorical and numerical.

A tree-based method involves the creation of a tree which progressively splits the observations into two groups, with the splits based on levels of the predictors, in order to group observations at each node (i.e. the tip of each branch) which have a similar outcome. Trees with a binary outcome are called decision trees. A good decision tree will group observations into nodes according to which of the two levels they take, with those observations at each node being the same level as much as possible. Trees with numerical outcomes are called regression trees and will group observations into nodes in order to minimise the variation in the outcome at each node, according to some defined criterion. Figure A.1 is an example of such a tree based on a numerical predictor,  $X_1$ , and a categorical predictor,  $X_2$ . There are five nodes ( $R_1, \dots, R_5$ ).



**Figure A.1:** An example of a tree.

The choice of splits can be optimised automatically in software to best separate those with high and low risk of seizures. The choices of splits yield useful information about the importance of the predictors. For example, for a categorical predictor, it indicates which levels are related to the outcome, and for numerical predictors, it can indicate cut-offs that are related to the outcome.

This approach is particularly flexible because it doesn't make any assumptions about the nature of the relationship between the predictors and the outcomes. For example,  $X_1$  needn't be linearly related to the outcome. This approach also allows for complex relationships between predictors by controlling



the depth or number of branches in a tree. For example, according to Figure A.1,  $X_2 = C$  is only useful as a predictor for those observations for which  $X_1 \leq t_2$ .

This combination of complexity and flexibility is important in exploratory situations, where there are no existing assumptions about the nature of the relationships between predictors and outcomes. This approach is also robust to outliers (that is, it is not overly influenced by a single observation) and can handle missing data by including this as a level of the predictor. Both of these issues can be relevant in the kinds of data available for these sub-projects.

In the context of data mining, there are typically a very large number of predictors available to predict each outcome. Hence it could be possible to provide a tree that is sufficiently complex to separate every observation into its own node. This may satisfy our criterion for similarity of the observations at each node, but such models are known to perform poorly on data which were not used in the construction of the model. Developing such an overspecified model is known as overfitting and needs to be actively avoided when the number of predictors is large.

If we want to avoid spurious relationships or generalise our model to wider contexts, we can avoid this problem with the creation of many trees based on random samples of the observations and averaging the results of these models. These kinds of models are called random forests, as they involve many trees generated from random samples of the data. These are now considered to be superior to classical methods in such contexts (Fernández-Delgado et al., 2014).

Boosting is an additional tool that can be applied to random forest models, particularly useful in the case of decision trees. Each new tree is generated according to weights so that those observations which were not classified unequivocally in the previous step have a greater weight than those that were, to improve prediction. These approaches are computationally demanding, but can be easily implemented in software such as R.

As these approaches involve the averaging of many tree models, there is no single tree that can be used to display the relationship between the predictors and the outcome. However, it is possible to consider the effect of each predictor, averaged over the others, using overall measures of importance. These measures of importance are the relative contribution of each predictor to all the trees in the forest. These are calculated by comparing the predictive power of trees with this predictor, compared to those with this predictor randomly permuted to distort its relationship to the outcome. It is also possible to visualise the relationship between the levels of the predictor and the outcome, averaged over the others, using partial plots. Examples of these are available for individual sub-projects.

Hastie et al. (2009) is an excellent resource for further reading on this topic.

## Appendix B

# Spatial Analysis of International Mail Interceptions

### B.1 modellingSA2.R

```
library(randomForest)
library(foreign)
library(rfPermute)

#####
##data preparation##
#####

#seizure data

#combining files
data1.1<-read.dbf("data_Sep14/Geocoding_Result_2.dbf")
data1.2<-read.dbf("data_Sep14/GC06AUG14.dbf")
year1.1<-as.Date(data1.1$inspection,format="%d/%m/%Y")
year1.2<-as.Date(data1.2$inspection,format="%d/%m/%y")
year1.1<-format(year1.1,format="%Y")
year1.2<-format(year1.2,format="%Y")
data1.1<-data1.1[year1.1>2007,]
data1.2<-data1.2[year1.2<2008,]

data1<-rbind(data1.1,data1.2)

#selecting just 2008 to 2012 for modelling
data1<-data1.1

data1$Comp_score<-NULL
data1$furtherdet<-NULL

#excluding no matches
data1<-data1[data1$SA1_2011>0,]

#excluding ties
data1<-data1[data1$Status!="M",]

#excluding declared
data1<-data1[data1$declaratio!="Declared",]

#obtaining consistent SA2 codes
data1$SA2_2011<-as.numeric(
```

```

    substring(as.character(data1$SA1_2011), 1,9) )

#storing seizures
write.csv(data1,"allseizures.csv")

#abs data

data2<-read.csv("ABS_all_SA2.csv")

#####
###merger-SA2 level###
#####

#collapse to one value for seizure id
temp<-model.matrix(~0+data1$SA2_2011+data1$seizureid_)
temp<-temp[,1]

#summarising seizure counts per SA2
data1b<-data.frame(table(temp))
colnames(data1b)<-c("SA2_2011","seizure.count")
data3b<-merge(data2,data1b,by=("SA2_2011"),
  all.x=TRUE, sort=TRUE)

#putting zeros in for SA2 without seizures
data3b$seizure.count[is.na(data3b$seizure.count)]<-0
#write.csv(data3b,"total_seizure_SA2.csv")

#column per category
#collapse to one value for seizure id by category
temp<-model.matrix(~0+data1$SA2_2011+
  data1$seizureid_+data1$category)
temp<-temp[,-2]

#summarising seizure counts per SA2 by category
n<-dim(temp)[2]

data1c<-names(table(temp[,1]))
for(i in 2:n)
{
  temp2<-table(temp[,1],temp[,i])[,2]
  data1c<-data.frame(data1c,temp2)
}
colnames(data1c)<-c("SA2_2011", levels(data1$category))

data3c<-merge(data3b,data1c,by=("SA2_2011"),
  all.x=TRUE, sort=TRUE)

data3c[is.na(data3c)]<-0

#summarising seizure counts per SA2 by specific commodities of interest
temp<-data.frame(table(data1$commodity))
temp<-temp[temp$Freq>1000,]
com<-temp[,1]

data1d<-data1[data1$commodity %in% com,]
data1d$commodity<-factor(data1d$commodity)
temp<-model.matrix(~0+data1d$SA2_2011+
  data1d$seizureid_+data1d$commodity)

```

```

temp<-temp[,-2]

#summarising seizure counts per SA2 by category
n<-dim(temp)[2]

data2d<-names(table(temp[,1]))
for(i in 2:n)
{
temp2<-table(temp[,1],temp[,i])[,2]
data2d<-data.frame(data2d,temp2)
}
colnames(data2d)<-c("SA2_2011", levels(data1d$commodity))

data2d[is.na(data2d)]<-0

#final data file

SA2full<-merge(data3c,data2d,by=("SA2_2011"),
  all.x=TRUE, sort=TRUE)
SA2full[is.na(SA2full)]<-0
rm(data1,data2,data1b,data1c,data3b,
  data1d, data2d, data3c, temp,temp2)

#####
###merger-SA2 and yearlevel###
#####

data1$year1<-as.numeric(data1$year1)
#collapse to one value for seizure id
temp<-model.matrix(~0+data1$SA2_2011+
data1$year1+data1$seizureid_)
temp<-temp[,1:2]

#summarising seizure counts per SA2
data1b<-data.frame(xtabs(~temp[,1]+temp[,2]))
colnames(data1b)<-c("SA2_2011","year","seizure.count")
data3b<-merge(data2,data1b,by=("SA2_2011"),
  all.x=TRUE, sort=TRUE)

#putting zeros in for levels without seizures
data3b$seizure.count[is.na(data3b$seizure.count)]<-0
#write.csv(data3b,"total_seizure_SA2.csv")

#column per category
#collapse to one value for seizure id by category
temp<-model.matrix(~0+data1$SA2_2011+
data1$year1+data1$seizureid_+data1$category)
temp<-temp[,-3]

#summarising seizure counts per SA2 and year by category
n<-dim(temp)[2]

temp2<-xtabs(~temp[,1]+temp[,2])
temp3<-as.data.frame.table(temp2)
data1c<-temp3[,1:2]
for(i in 3:n)
{

```

```

temp2<-xtabs(~temp[,1]+temp[,2]+temp[,i])[,2]
temp3<-as.data.frame.table(temp2)[,3]
data1c<-data.frame(data1c,temp3)
}
colnames(data1c)<-c("SA2_2011",
"year",levels(data1$category))

data3c<-merge(data3b,data1c,
by=c("SA2_2011", "year"), all.x=TRUE, sort=TRUE)

data3c[is.na(data3c)]<-0

#summarising seizure counts per SA2 and year
#by specific commodities of interest
com<-c("Apple","Beef","Cacti/Succulents","Khat","Mooncakes with egg",
"Mooncakes with meat","Pork","Shamrock","Tea")

#collapse to one value for seizure id by commodity
data1d<-data1[data1$commodity %in% com,]
data1d$commodity<-factor(data1d$commodity)
temp<-model.matrix(~0+data1d$SA2_2011+
data1d$year1+data1d$seizureid_+data1d$commodity)
temp<-temp[,-3]

n<-dim(temp)[2]
temp2<-xtabs(~temp[,1]+temp[,2])
temp3<-as.data.frame.table(temp2)
data2d<-temp3[,1:2]
for(i in 3:n)
{
temp2<-xtabs(~temp[,1]+temp[,2]+temp[,i])[,2]
temp3<-as.data.frame.table(temp2)[,3]
data2d<-data.frame(data2d,temp3)
}
colnames(data2d)<-c("SA2_2011", "year",
levels(data1d$commodity))

data2d[is.na(data2d)]<-0

SA2yearfull<-merge(data3c,data2d,
by=c("SA2_2011", "year"), all.x=TRUE, sort=TRUE)
SA2yearfull[is.na(SA2yearfull)]<-0
write.csv(SA2yearfull,"SA2full_year.csv")

rm(data1,data2,data1b,data1c,data3b,
data1d, data2d, data3c, temp,temp2)

#####
###Analysis###
#####

#random forests
vars<-colnames(SA2full[,3:109])

outtotal<-"seizure.count"

#if rate per 100,000 used

```

```

#SA2full$myrate <- #SA2full$seizure.count/#SA2full$Total_persons*100000
#SA2full$myrate[#SA2full$Total_persons < 1000] <- 0
#outtotal<-"myrate"
#also change the figure y axis to rate per 100,000 not total

outcat<-colnames(SA2full[,111:123])
outcom<-colnames(SA2full[,124:(123+length(com))])

#total
fvars<-as.formula(paste(paste(outtotal,"~",sep=""),
                        paste(vars,collapse="+")))
set.seed(56)
forest1<-randomForest(fvars,data=SA2full,ntree=500)
set.seed(56)
forest1p<-rfPermute(fvars,data=SA2full,ntree=500,nrep=100)
imp<-importance(forest1)/max(importance(forest1))*100
import1<-round(data.frame(imp,forest1p$null.dist$pval[,2]),4)
import1<-import1[order(import1[,1],decreasing=T),]
import1<-import1/import1[1]*100
colnames(import1)<-c("importance","p-value")
write.csv(round(import1,4),"rf_importance_SA2_rate.csv")
impvars<-rownames(import1[1:12,])

pdf("partialPlot_Total.pdf", width=6,height=9)
par(mfrow=c(3,2),oma = c(0,0,0,0) + 0.1,
    mar = c(4,4,1,1) + 0.1)
for(i in 1:12)
{
  temp<-partialPlot(forest1,x.var=impvars[i], SA2full,
                    ylab="Predicted seizure count\n (all seizures)",
                    main="", xlab=impvars[i])
  text(max(temp$x)-0.25*(max(temp$x)-min(temp$x)),
       min(temp$y)+0.1*(max(temp$y)-min(temp$y)),
       paste("Relative importance=\n",round(import1[i,1],2),
            "\n P-value=",round(import1[i,2],3),sep=""))
}
dev.off()

#broad categories
import2<-vars
for(j in 1:length(outcat))
{
  fvars<-as.formula(paste(paste("'",outcat[j],"'",sep=""),
                          paste(vars,collapse="+")))
  set.seed(56)
  forest1<-randomForest(fvars,data=SA2full,ntree=500)
  set.seed(56)
  forest1p<-rfPermute(fvars,data=SA2full,ntree=500,nrep=50)
  imp<-importance(forest1)/max(importance(forest1))*100
  import1<-data.frame(imp,forest1p$null.dist$pval[,2])
  colnames(import1)<-c(paste(outcat[j],"-importance",sep=""),
                      paste(outcat[j],"-p-value",sep=""))
  import2<-data.frame(import2,import1)
  import1<-import1[
    order(import1[,1],decreasing=T),]
  impvars<-rownames(import1[1:12,])
}

```

```

pdf(paste("partialPlot_categories/",gsub("/", "_",
      paste("partialPlot1_",outcat[j],".pdf",sep="")),sep=""),
    width=6,height=9)
par(mfrow=c(3,2),oma = c(0,0,0,0) + 0.1,
    mar = c(4,4,1,1) + 0.1)
for(i in 1:12)
{
  temp<-partialPlot(forest1,x.var=impvars[i], SA2full,
    ylab=paste("Predicted count\n (",outcat[j],")",sep=""),
    main="", xlab=impvars[i])
  text(max(temp$x)-0.25*(max(temp$x)-min(temp$x)),
    min(temp$y)+0.1*(max(temp$y)-min(temp$y)),
    paste("Relative importance=\n",round(import1[i,1],2),
      "\n P-value=",round(import1[i,2],3),sep=""))
}
dev.off()

}
write.csv(import2,"rf_importance_SA2_total_cat.csv")

#key commodities
import3<-vars
for(j in 1:length(outcom))
{
  fvars<-as.formula(paste(paste("'",outcom[j],"'~",sep=""),
    paste(vars,collapse="+"))))
  set.seed(56)
  forest1<-randomForest(fvars,data=SA2full,ntree=500)
  set.seed(56)
  forest1p<-rfPermute(fvars,data=SA2full,ntree=500,nrep=50)
  imp<-importance(forest1)/max(importance(forest1))*100
  import1<-data.frame(imp,forest1p$null.dist$pval[,2])
  colnames(import1)<-c(paste(outcom[j],"-importance",sep=""),
    paste(outcom[j],"-p-value",sep=""))
  import3<-data.frame(import3,import1)
  import1<-import1[
    order(import1[,1],decreasing=T),]
  impvars<-rownames(import1[1:12,])

  pdf(paste("partialPlot_keycommodities/",gsub("/", "_",
    paste("partialPlot1_",outcom[j],".pdf",sep="")),
    sep=""), width=6,height=9)
  par(mfrow=c(3,2),oma = c(0,0,0,0) + 0.1,
    mar = c(4,4,1,1) + 0.1)
  for(i in 1:12)
  {
    temp<-partialPlot(forest1,x.var=impvars[i], SA2full,
      ylab=paste("Predicted count\n (",outcom[j],")",sep=""),
      main="", xlab=impvars[i])
    text(max(temp$x)-0.25*(max(temp$x)-min(temp$x)),min(temp$y)+
      0.1*(max(temp$y)-min(temp$y)),
      paste("Relative importance=\n",round(import1[i,1],2),
        "\n P-value=",round(import1[i,2],3),sep=""))
  }
  dev.off()
}
write.csv(import3,"rf_importance_SA2_total_com.csv")

```

```
#####
##languages only##
#####

#key languages only
lang<-SA2full[,c(17:18,20:77)] #all languages except English and not stated
vars<-names(lang)
outtotal<-"seizure.count"
fvars<-as.formula(paste(paste(outtotal,"~",sep=""),
  paste(vars,collapse="+"))))
set.seed(56)
forest1<-randomForest(fvars,data=SA2full,ntree=500)
set.seed(56)
forest1p<-rfPermute(fvars,data=SA2full,ntree=500,nrep=100)
imp<-importance(forest1)/max(importance(forest1))*100
import1<-data.frame(imp,forest1p$null.dist$pval[,2])
import1<-import1[order(import1[,1],decreasing=T),]
colnames(import1)<-c("importance","p-value")
write.csv(round(import1,4),"rf_importance_SA2_total_language.csv")
impvars<-rownames(import1)
maxy<-maxx<-0
for(i in 1:5)
{
temp<-partialPlot(forest1,x.var=impvars[i], SA2full,
  ylab="Predicted seizure count (Total)", main="")
maxy<-max(temp$y,maxy)
maxx<-max(temp$x,maxx)
}
pdf("partialPlot_Total_language.pdf")
for(i in 1:5)
{
partialPlot(forest1,x.var=impvars[i], SA2full,
  ylab="Predicted seizure count (Total)", main="",
  xlab="Number of speakers", xlim=c(0,maxx),
  ylim=c(0,maxy), col=i, rug=F)
par(new=T)
}
Language<-gsub("_"," ",gsub("Language_","",paste(impvars[1:5],
  "(",round(import1[1:5,1],1),",", "(",round(import1[1:5,2],2),")",
  sep="")))
legend("bottomright",col=1:5,Language,lty=1, title="Language (importance, p-value)")
dev.off()

#broad categories
import2<-vars
for(j in 1:length(outcat))
{
fvars<-as.formula(paste(paste("'",outcat[j],"'",sep=""),
  paste(vars,collapse="+"))))
set.seed(56)
forest1<-randomForest(fvars,data=SA2full,ntree=500)
set.seed(56)
forest1p<-rfPermute(fvars,data=SA2full,ntree=500,nrep=50)
imp<-importance(forest1)/max(importance(forest1))*100
import1<-data.frame(imp,forest1p$null.dist$pval[,2])
colnames(import1)<-c(paste(outcat[j],"-importance",sep=""),
```



```

        paste(outcat[j], "-p-value", sep=""))
import2<-data.frame(import2,import1)
import1<-import1[
  order(import1[,1],decreasing=T),]
impvars<-rownames(import1)

maxy<-maxx<-0
for(i in 1:5)
{
  temp<-partialPlot(forest1,x.var=impvars[i], SA2full,
    ylab="", main="")
  maxy<-max(temp$y,maxy)
  maxx<-max(temp$x,maxx)
}

pdf(paste("partialPlot_categories/",gsub("/","_",
  paste("partialPlot1_",outcat[j],"_language.pdf",sep="")),sep=""),
  width=6,height=9)

for(i in 1:5)
{
  partialPlot(forest1,x.var=impvars[i], SA2full,
    ylab=paste("Predicted count\n (",outcat[j],")",sep=""),
    main="", xlab="Number of speakers", xlim=c(0,maxx),
    ylim=c(0,maxy), col=i, rug=F)
  par(new=T)
}
Language<-gsub("_"," ",gsub("Language_","",paste(impvars[1:5],
" (",round(import1[1:5,1],1),", ", round(import1[1:5,2],2),")",
sep="")))
legend("bottomright",col=1:5,Language,lty=1, title="Language (importance, p-value)")
dev.off()
}
write.csv(import2,"rf_importance_SA2_total_cat_language.csv")

#key commodities
import3<-vars
for(j in 1:length(outcom))
{
  fvars<-as.formula(paste(paste("'",outcom[j],"'~",sep=""),
    paste(vars,collapse="+")))
  set.seed(56)
  forest1<-randomForest(fvars,data=SA2full,ntree=500)
  set.seed(56)
  forest1p<-rfPermute(fvars,data=SA2full,ntree=500,nrep=50)
  imp<-importance(forest1)/max(importance(forest1))*100
  import1<-data.frame(imp,forest1p$null.dist$pval[,2])
  colnames(import1)<-c(paste(outcom[j],"-importance",sep=""),
    paste(outcom[j],"-p-value",sep=""))
  import3<-data.frame(import3,import1)
  import1<-import1[
    order(import1[,1],decreasing=T),]
  impvars<-rownames(import1)

  maxy<-maxx<-0
  for(i in 1:5)
  {

```

```

temp<-partialPlot(forest1,x.var=impvars[i], SA2full,
  ylab="", main="")
maxy<-max(temp$y,maxy)
maxx<-max(temp$x,maxx)
}

pdf(paste("partialPlot_keycommodities/",gsub("/", "_",
  paste("partialPlot1_",outcom[j],"_language.pdf",sep="")),sep=""),
  width=6,height=9)

for(i in 1:5)
{
partialPlot(forest1,x.var=impvars[i], SA2full,
  ylab=paste("Predicted count\n (" ,outcom[j],")",sep=""),
  main="", xlab="Number of speakers", xlim=c(0,maxx),
  ylim=c(0,maxy), col=i, rug=F)
par(new=T)
}
Language<-gsub("_"," ",gsub("Language_","",paste(impvars[1:5],
" (",round(import1[1:5,1],1),", " ,round(import1[1:5,2],2),")",
sep="")))
legend("bottomright",col=1:5,Language,lty=1, title="Language (importance, p-value)")
dev.off()
}
write.csv(import3,"rf_importance_SA2_total_com_language.csv")

```

## B.2 spatialplots.R

```
#note this script assumes the structure:
#all input data are in a folder "indat"
#all outputs are stored in a folder "output"

library(ggmap)
library(rgdal)
library(gridExtra)

# readin data
sa2dat <- read.csv("../indat/SA2full_year.csv")
sa2dat[sa2dat$Total_persons == 0, c("SA2_MAIN11",
"SA2_NAME11", "year", "seizure.count", "Total_persons")]

# Calculate seizure rate per 100,000 persons
sa2dat$myrate <- sa2dat$seizure.count/sa2dat$Total_persons*100000
# Some of the total populations are very small causing
# erroneously high rates for these SA2s
sa2dat$myrate[sa2dat$Total_persons < 1000] <- 0
summary(sa2dat$myrate)
sa2dat[sa2dat$myrate > 1000, c("SA2_MAIN11", "SA2_NAME11",
"year", "seizure.count", "Total_persons", "myrate")]

# readin the raw seizures data
tmp <- read.csv("../indat/allseizures.csv")
allseizures <- subset(tmp, select=c("X", "Y",
"year1", "category", "subcategor", "commodity", "SA2_2011"))
colnames(allseizures) <- c("lon", "lat", "year", "category",
"subcategor", "commodity", "SA2_MAIN11")

#####
# read ABS file containing AGCS data structures #
#####

#abs files obtained from
#http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202011

tmp <- read.csv("../indat/MB_2011_VIC.csv")
abs <- unique(tmp)
# Select out Greater Melbourne
gm_abs <- subset(abs, GCC_NAME11 == "Greater Melbourne")

#####
##### CONTOUR PLOTS WITH RAW DATA #####
#####

# select out Greater Melbourne
dat <- merge(allseizures, gm_abs, by="SA2_MAIN11", all.y = TRUE)
mdat <- subset(dat, subset=(year == 2008 | year == 2009 |
year == 2010 | year == 2011))

MelbourneMap <- qmap(c(lon=145.15, lat=-37.81411), zoom=9)

pdf(file="../output/Melb_rawdat.pdf", width = 10, height = 10)
MelbourneMap + geom_point(aes(x = lon, y = lat, colour = "red"),
data = mdat) + labs(title="All Seizures during 2008-2011") +
```

```

    theme(title=element_text(size=9, face="bold"), legend.position="none")
dev.off()

pdf(file="./output/Melb_rawdat_byyear.pdf",width = 10, height = 10)
MelbourneMap + geom_point(aes(x = lon, y = lat, colour = "red"),
  data = mdat) + labs(title="All Seizures during 2008-2011") +
  theme(title=element_text(size=9, face="bold"),
    legend.position="none") + facet_wrap(~ year)
dev.off()

# contour plot
pdf(file="./output/Melb_contour.pdf",width = 10, height = 10)
MelbourneMap + stat_density2d(geom="polygon",
  aes(fill = ..level..), bins = 200, contour = TRUE,
  data = mdat) + scale_fill_gradient(low = "black",high= "red")
dev.off()

# panel contour by year
pdf(file="./output/Melb_contour_byyear.pdf",width = 10, height = 10)
MelbourneMap + stat_density2d(geom="polygon",
  aes(fill = ..level..), bins = 200, contour = TRUE, data = mdat) +
  scale_fill_gradient(low = "black",high= "red") + facet_wrap(~ year)
dev.off()

#by key commodity

com<-c("Apple","Beef","Cacti/Succulents","Khat",
"Mooncakes with egg","Mooncakes with meat","Pork",
"Shamrock","Tea")
comcode<- c("Apple","Beef","Cacti","Khat","Mooncakes_egg",
"Mooncakes_meat","Pork","Shamrock","Tea")
n<-length(com)
for(i in 1:9){
mdatcom<-mdat[mdat$commodity==com[i],]
mdatcom<-mdatcom[is.na(mdatcom$commodity)==F,]
pdf(file=paste("./output/Melb_",comcode[i],"_rawdat_byyear.pdf",sep=""),
  width = 10, height = 10)
MelbourneMap + geom_point(aes(x = lon, y = lat, colour = "red"),
  data = mdatcom) + labs(title=paste(com," Seizures during 2008-2011",sep="")) +
  theme(title=element_text(size=9, face="bold"),
    legend.position="none") + facet_wrap(~ year)
dev.off()
}

#####
### SA2 level analysis ##
#####

# Restrict to most recent 4 years of data
dat_long <- subset(sa2dat, subset=(year == 2008 |
  year == 2009 | year == 2010 | year == 2011),
  select=c("SA2_MAIN11","SA2_NAME11","Total_persons",
    "year","seizure.count","myrate"))
dat_wide <- reshape(dat_long, timevar="year",
  idvar = c("SA2_MAIN11","SA2_NAME11"), direction = 'wide')

# Merge data sources (restricting to Greater Melbourne)
mdat <- merge(dat_wide, gm_abs, by="SA2_MAIN11", all.y = TRUE)

```

```

# Replace NA with 0
mdat$seizure.count.2008[is.na(mdat$seizure.count.2008)] <- 0
mdat$seizure.count.2009[is.na(mdat$seizure.count.2009)] <- 0
mdat$seizure.count.2010[is.na(mdat$seizure.count.2010)] <- 0
mdat$seizure.count.2011[is.na(mdat$seizure.count.2011)] <- 0
mdat$myrate.2008[is.na(mdat$myrate.2008) | mdat$myrate.2008 == 0] <- 0.01
mdat$myrate.2009[is.na(mdat$myrate.2009) | mdat$myrate.2008 == 0] <- 0.01
mdat$myrate.2010[is.na(mdat$myrate.2010) | mdat$myrate.2008 == 0] <- 0.01
mdat$myrate.2011[is.na(mdat$myrate.2011) | mdat$myrate.2008 == 0] <- 0.01

#abs files obtained from
#http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202011
aus.sa2 <- readOGR("./indat","SA2_2011_AUST")

# select out Greater Melbourne area
gm_sa2 <- aus.sa2[aus.sa2$GCC_NAME11 == "Greater Melbourne",]
rlabs <- row.names(gm_sa2)

# merge study data into the shapefile
gm_sa2@data <- merge(x = gm_sa2@data, y = mdat, by="SA2_MAIN11", all.x = TRUE)
row.names(gm_sa2@data) <- rlabs
#gm_sa2@data[,c("SA2_NAME11","seizure.count.2008")]

print(proj4string(gm_sa2))

# convert shapefile into format compatible with qmap
data <- fortify(gm_sa2)
# need to remerge study data back into this new format
data <- merge(data, gm_sa2@data, by.x="id", by.y="row.names")

#####
#### FIGURES: COUNTS ####
#####

summary(mdat$seizure.count.2008)
summary(mdat$seizure.count.2009)
summary(mdat$seizure.count.2010)
summary(mdat$seizure.count.2011)

pdf(file="../output/Melb_sa2_ex1_counts.pdf",width = 10, height = 10)
p1 <- ggplot(gm_sa2) + geom_polygon(aes(x = long, y = lat,
  group = group, fill = seizure.count.2008), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,400), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Counts in 2008") +
  theme(title=element_text(size=9, face="bold"), legend.position="right")
p2 <- ggplot(gm_sa2) + geom_polygon(aes(x = long, y = lat,
  group = group, fill = seizure.count.2009), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,400), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Counts in 2009") +
  theme(title=element_text(size=9, face="bold"), legend.position="right")
p3 <- ggplot(gm_sa2) + geom_polygon(aes(x = long, y = lat,
  group = group, fill = seizure.count.2010), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,400), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates Counts in 2010") +

```

```

      theme(title=element_text(size=9, face="bold"), legend.position="right")
p4 <- ggplot(gm_sa2) + geom_polygon(aes(x = long, y = lat,
  group = group, fill = seizure.count.2011), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,400), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates Counts in 2011") +
  theme(title=element_text(size=9, face="bold"), legend.position="right")
grid.arrange(p1,p2,p3,p4, heights=1:1, widths=1:1)
dev.off()

```

```

MelbourneMap <- qmap(c(lon=145.15,lat=-37.81411),zoom=9)

```

```

pdf(file="../output/Melb_sa2_ex2_counts.pdf",width = 10, height = 10)
p1 <- MelbourneMap + geom_polygon(aes(x = long, y = lat,
  group = group, fill = seizure.count.2008), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,400), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates Counts in 2008") +
  theme(title=element_text(size=9, face="bold"), legend.position="right")
p2 <- MelbourneMap + geom_polygon(aes(x = long, y = lat,
  group = group, fill = seizure.count.2009), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,400), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates Counts in 2009") +
  theme(title=element_text(size=9, face="bold"), legend.position="right")
p3 <- MelbourneMap + geom_polygon(aes(x = long, y = lat,
  group = group, fill = seizure.count.2010), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,400), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates Counts in 2010") +
  theme(title=element_text(size=9, face="bold"), legend.position="right")
p4 <- MelbourneMap + geom_polygon(aes(x = long, y = lat,
  group = group, fill = seizure.count.2011), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,400), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates Counts in 2011") +
  theme(title=element_text(size=9, face="bold"), legend.position="right")
grid.arrange(p1,p2,p3,p4, heights=1:1, widths=1:1)
dev.off()

```

```

#####
#### FIGURES: RATES ####
#####

```

```

mdat[,c("SA2_MAIN11","SA2_NAME11.x","Total_persons.2008",
"seizure.count.2008","myrate.2008")]

```

```

summary(mdat$myrate.2008)
summary(mdat$myrate.2009)
summary(mdat$myrate.2010)
summary(mdat$myrate.2011)

```

```

pdf(file="../output/Melb_sa2_ex1_rates.pdf",width = 10, height = 10)
p1 <- ggplot(gm_sa2) + geom_polygon(aes(x = long, y = lat,
  group = group, fill = myrate.2008), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,1700), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates (per 100,000 person) in 2008") +

```

```

      theme(title=element_text(size=9, face="bold"), legend.position="none")
p2 <- ggplot(gm_sa2) + geom_polygon(aes(x = long, y = lat,
  group = group, fill = myrate.2009), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,1700), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates (per 100,000 person) in 2009") +
  theme(title=element_text(size=9, face="bold"), legend.position="none")
p3 <- ggplot(gm_sa2) + geom_polygon(aes(x = long, y = lat,
  group = group, fill = myrate.2010), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,1700), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates (per 100,000 person) in 2010") +
  theme(title=element_text(size=9, face="bold"), legend.position="none")
p4 <- ggplot(gm_sa2) + geom_polygon(aes(x = long, y = lat,
  group = group, fill = myrate.2011), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,1700), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates (per 100,000 person) in 2011") +
  theme(title=element_text(size=9, face="bold"), legend.position="none")
grid.arrange(p1,p2,p3,p4, heights=1:1, widths=1:1)
dev.off()

MelbourneMap <- qmap(c(lon=145.15,lat=-37.81411),zoom=9)
pdf(file="../../output/Melb_sa2_ex2_rates.pdf",width = 10, height = 10)
p1 <- MelbourneMap + geom_polygon(aes(x = long, y = lat,
  group = group, fill = myrate.2008), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,1700), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates (per 100,000 person) in 2008") +
  theme(title=element_text(size=9, face="bold"), legend.position="right")
p2 <- MelbourneMap + geom_polygon(aes(x = long, y = lat,
  group = group, fill = myrate.2009), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,1700), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates (per 100,000 person) in 2009") +
  theme(title=element_text(size=9, face="bold"), legend.position="right")
p3 <- MelbourneMap + geom_polygon(aes(x = long, y = lat,
  group = group, fill = myrate.2010), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,1700), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates (per 100,000 person) in 2010") +
  theme(title=element_text(size=9, face="bold"), legend.position="right")
p4 <- MelbourneMap + geom_polygon(aes(x = long, y = lat,
  group = group, fill = myrate.2011), data = data,
  alpha = .7, size = .3) + scale_fill_gradient(limits=c(0,1700), name="seizures",
  trans='sqrt',low = "black", high= "red") +
  labs(title="Seizure Rates (per 100,000 person) in 2011") +
  theme(title=element_text(size=9, face="bold"), legend.position="right")
grid.arrange(p1,p2,p3,p4, heights=1:1, widths=1:1)
dev.off()

```

## Appendix C

# Generalised Pattern Analysis for International Passengers

```
#####
#### data preparation ####
#####
load("dat2_v2.rdata")
load("dat_v2.rdata")
dat2a<-dat2[dat2$Local_Port_Code=="ADL",]
load("SYD_data.rdata")
names(SYD_data)[11]<-"Local_Port_Code2"

levels(SYD_data$Seizure)<-c(1,1,0)
levels(dat2a$Seizure)<-c(0,1)
dat2<-rbind(dat2a,SYD_data)

#collapsing endpoints assuming action
dat2$Endpoint2<-dat2$Endpoint
dat2$Endpoint2[dat2$Endpoint=="3a. DEC Xray not done/not recorded"] <-
"3. DEC Xray only"
dat2$Endpoint2[dat2$Endpoint=="14a. NONDEC Xray not done/not recorded"] <-
"14. NONDEC Xray only"
dat2$Endpoint2[dat2$Endpoint=="6a. DEC K9 not done/not recorded"] <-
"6. DEC K9 only"
dat2$Endpoint2[dat2$Endpoint=="16a. NONDEC K9 not done/not recorded"] <-
"16. NONDEC K9 only"
dat2$Endpoint2[dat2$Endpoint=="4. DEC Xray + dec insp"] <-
"4-5. DEC Xray + insp"

#creating summaries of endpoints
dat2$Screening<-dat2$Declaration<-dat2$Inspection<-dat2$Endpoint

levels(dat2$Screening)[levels(dat2$Screening) %in%
levels(dat2$Endpoint)[c(1,5,6,7,16)]]<-"None"
levels(dat2$Screening)[levels(dat2$Screening) %in%
levels(dat2$Endpoint)[c(8,9,10,17,18,19,20,21,28)]]<-"Xray"
levels(dat2$Screening)[levels(dat2$Screening) %in%
levels(dat2$Endpoint)[c(11,12,13,22,23,24,25)]]<-"K9"
levels(dat2$Screening)[levels(dat2$Screening) %in%
levels(dat2$Endpoint)[c(2,3,4,14,15,26)]]<-"Manual"

levels(dat2$Declaration)[levels(dat2$Declaration) %in%
levels(dat2$Endpoint)[5:15]]<-"Declarant"
```



```

levels(dat2$Declaration)[levels(dat2$Declaration) %in%
levels(dat2$Endpoint)[c(1:4,16:26)]]<-"Non-declarant"

levels(dat2$Inspection)[levels(dat2$Inspection) %in%
levels(dat2$Endpoint)[c(1,16:18,22,23,24,26,5:9,11,12,14)]]<-"N/A"
levels(dat2$Inspection)[levels(dat2$Inspection) %in%
levels(dat2$Endpoint)[c(19,20,25,2,3)]]<-"Compliance"
levels(dat2$Inspection)[levels(dat2$Inspection) %in%
levels(dat2$Endpoint)[c(21,4,10,13,15)]]<-"Non-compliance"

table(dat2$Screening, dat2$Inspection)
table(dat2$Declaration)
prop.table(table(dat2$Screening, dat2$Seizure),1)

screens<-levels(dat2$Screening)
dat2$Seizure<-as.numeric(dat2$Seizure)
dat2$Seizure<-dat2$Seizure-1

```

## Appendix D

# Detecting Anomalous Broker Activity

```
#####
#### importing data ####
#####

require(tools)

# Test
options(stringsAsFactors=FALSE)
dat <- read.delim("3. AIMS_Directions.tab")
str(dat)
# Date format is: 14/01/2013 9:49
getdate <- function(x) as.POSIXct(strptime(x, format="%d/%m/%Y %H:%M"))
x <- getdate(dat$Initiatingdate)
head(x)

# Set nrows to, say, 10 for testing, 0 for production
nrows <- 0
(tabfiles <- list.files(pattern=glob2rx("*.tab")))
(datanames <- make.names(substr(tabfiles, 4, nchar(tabfiles)-4)))
(datanames <- gsub(".", "_", datanames, fixed=TRUE))
for (i in 1:length(tabfiles)) {
  dat <- read.delim(tabfiles[i], nrows=nrows)
  datecolumns <- grep("Msg_Id|date", names(dat))
  for (j in datecolumns) dat[,j] <- getdate(dat[,j])
  assign(datanames[i], dat, envir=.GlobalEnv)
}
save.image("data.RData")

#####
#### combining data ####
#####

# combining CP files

#add missing column
CEBRA_decs_with_CP_Apr__Jun_13$CP_Risk_Type<-rep(NA,dim(CEBRA_decs_with_CP_Apr__Jun_13)[1])
CEBRA_decs_with_CP_Jan_Mar_13$CP_Risk_Type<-rep(NA,dim(CEBRA_decs_with_CP_Jan_Mar_13)[1])
CEBRA_decs_with_CP_Oct_Dec_13$CP_Risk_Type<-rep(NA,dim(CEBRA_decs_with_CP_Oct_Dec_13)[1])
CEBRA_decs_with_CP_Aug_13$CP_Risk_Type<-rep(NA,dim(CEBRA_decs_with_CP_Aug_13)[1])
CP<-get(datanames[2])
for(i in 3:7)
```

```

{
  CP<-rbind(CP,get(datanames[i]))
}

#####
#####Threatening AIMS Directions#####
#####

ndir<-dim(AIMS_Directions)[1]
#indicating threatening directions or not
threats<-c("AIR Freight Inspection",
           "ICS amendment required - Food",
           "ICS amendment required - Qtine",
           "Inspect (Hold Seals Intact)",
           "Inspect (unpack)",
           "Nursery Stock Inspect",
           "SIP - Hold Seals Intact",
           "SIP - Inspect (Unpack)",
           "Under Surveillance",
           "Verify (Hold Seals Intact)",
           "Verify certs Bulk Timber Insp.",
           "Verify Commodity",
           "Verify Packing",
           "Verify prior to Man. Fum.",
           "Verify prior to VolFum.",
           "Verify Tarping",
           "Withdrawn Entry"
)
nthr<-length(threats)

for(i in 1:ndir)
{
  AIMS_Directions$threat[i]<-1*(AIMS_Directions$Direction[i] %in% threats)
}

#####
#amendments#
#####

#initialising
beforedate<-afterdate<-vector()
class(beforedate) <- class(afterdate) <- class(AIMS_Directions$Initiatingdate)
for(i in 1:ndir)
{
  id<-AIMS_Directions$Quarantine.Entry[i] #entry id for that direction
  dt<-AIMS_Directions$Initiatingdate[i] #date of the direction
  #selecting details for that direction
  temp<- subset(CEBRA_decs_with_line_details,
               CEBRA_decs_with_line_details$Declaration_Id==id)
  datesi<-unique(temp$FID_Transaction_Msg_Id)
  ndate<-length(datesi)
  if(ndate>1)
  {

    if(dt<datesi[ndate])
    {
      for(j in 1:(ndate-1))
      {

```

```

        gap1<-difftime(dt,datesi[j], unit="days")
        if((gap1>0)&(gap1<30)){beforedate[i]<-datesi[j]}
        if((gap1>0)&(gap1<30)){afterdate[i]<-datesi[j+1]}
    }
}
}
#diff1<-dim(AIMS_Directions)[1]-length(beforedate)
#beforedate1<-c(beforedate,rep(0,diff1))
#afterdate1<-c(afterdate,rep(0,diff1))
write.csv(data.frame(AIMS_Directions$Quarantine.Entry,
    AIMS_Directions$Initiatingdate,beforedate,afterdate),
    "AmendDates.csv")

#just a count of corresponding amendments
AmendAfter<-1*(afterdate>0)
AmendAfter[is.na(AmendAfter)] <- 0

#testing if threatening Directions result in more amendments:
fisher.test(table(AmendAfter,AIMS_Directions$threat))

#overall patterns of directions
tab1<-prop.table(table(AIMS_Directions$Direction,AmendAfter),1)
tab1<-cbind(tab1,table(AIMS_Directions$Direction))
tab1[order(tab1[,2],decreasing=TRUE),]
write.csv(tab1[order(tab1[,2],decreasing=TRUE),],"AmendbyDirection.csv")

#testing time gap
#gap1<-afterdate-AIMS_Directions$Initiatingdate

##breaking down by type of amendment
#STILL TO ADD:
#quantity
#high risk answer to CP

#just relevant rows
amendrows<-which(AmendAfter==1)

#look at the before and after date for each direction and which
diffDelivery_Address<-diffImporter_Id<-diffImporter_Address<-vector()
changeSupplier_Id<-changeImporter_Id<-changeBrokerage_Id<-diffTariff_Class<-vector()
diffGoods_Quantity_Value_1<-diffGoods_Description<-vector()
diffLine_Nbr<-addLine_Nbr<-deleteLine_Nbr<-vector()
for(i in amendrows)
{
    id<-AIMS_Directions$Quarantine.Entry[i] #entry id for that direction

    #getting line details and selecting two time points
    tempi<- subset(CEBRA_decs_with_line_details,
        CEBRA_decs_with_line_details$Declaration_Id==id)
    beforei<-subset(tempi,tempi$FID_Transaction_Msg_Id==beforedate[i])
    afteri<-subset(tempi,tempi$FID_Transaction_Msg_Id==afterdate[i])

    diffLine_Nbr[i]<-length(unique(afteri$Line_Nbr))-length(unique(beforei$Line_Nbr))
    addLine_Nbr[i]<-length(setdiff(unique(afteri$Line_Nbr),
        unique(beforei$Line_Nbr)))
    deleteLine_Nbr[i]<-length(setdiff(unique(beforei$Line_Nbr),

```

```

    unique(afteri$Line_Nbr)))
diffTariff_Class[i]<-length(setdiff(unique(beforei$Tariff_Class),
    unique(afteri$Tariff_Class)))
diffGoods_Description[i]<-length(setdiff(unique(beforei$Goods_Description),
    unique(afteri$Goods_Description)))
diffGoods_Quantity_Value_1[i]<-length(setdiff(unique(beforei$Quantity_Value_1),
    unique(afteri$Quantity_Value_1)))

#focusing on line 1 only for entry level characteristics
beforei1<-beforei[1,]
afteri1<-afteri[1,]

diffDelivery_Address[i]<-adist(afteri1$Delivery_Address,
    beforei1$Delivery_Address)
diffImporter_Address[i]<-adist(afteri1$Importer_Address,
    beforei1$Importer_Address)
changeImporter_Id[i]<-1-(afteri1$Importer_Id==beforei1$Importer_Id)
changeSupplier_Id[i]<-1-(afteri1$Supplier_Id==beforei1$Supplier_Id)
changeBrokerage_Id[i]<-1-(afteri1$Brokerage_Id==beforei1$Brokerage_Id)

}

addLineANY_Nbr<-1*(addLine_Nbr>0)
deleteLineANY_Nbr<-1*(deleteLine_Nbr>0)
diffLineANYPOS_Nbr<-1*(diffLine_Nbr>0)
diffLineANYNEG_Nbr<-1*(diffLine_Nbr<0)
diffDelivery_AddressANY<-1*(diffDelivery_Address>0)
diffImporter_AddressANY<-1*(diffImporter_Address>0)
diffTariff_ClassANY<-1*(diffTariff_Class>0)
diffGoods_Quantity_Value_1ANY<-1*(diffGoods_Quantity_Value_1>0)
diffGoods_DescriptionANY<-1*(diffGoods_Description>0)

fisher.test(table(changeImporter_Id,AIMS_Directions$threat))
fisher.test(table(changeSupplier_Id,AIMS_Directions$threat))
fisher.test(table(changeBrokerage_Id,AIMS_Directions$threat))
fisher.test(table(addLineANY_Nbr,AIMS_Directions$threat))
fisher.test(table(deleteLineANY_Nbr,AIMS_Directions$threat))
fisher.test(table(diffLineANYPOS_Nbr,AIMS_Directions$threat))
fisher.test(table(diffLineANYNEG_Nbr,AIMS_Directions$threat))
pdf("Delivery_boxplot.pdf",width=7, height=4)
boxplot(diffDelivery_Address[diffDelivery_Address>1]~
    AIMS_Directions$threat[diffDelivery_Address>1],
    yaxt="n",horizontal=TRUE,
    xlab="generalized Levenshtein distance",
    ylab="Threatening Direction")
axis(2,1:2,c("No","Yes"),las=1)
dev.off()
fisher.test(table(diffDelivery_AddressANY,AIMS_Directions$threat))

pdf("Importer_boxplot.pdf",width=7, height=4)
boxplot(diffImporter_Address[diffImporter_Address>1]~
    AIMS_Directions$threat[diffImporter_Address>1],
    yaxt="n",horizontal=TRUE,
    xlab="generalized Levenshtein distance",
    ylab="Threatening Direction")
axis(2,1:2,c("No","Yes"),las=1)
dev.off()
fisher.test(table(diffImporter_AddressANY,AIMS_Directions$threat))

```

```

pdf("Tariff_boxplot.pdf",width=7, height=4)
boxplot(diffTariff_Class[diffTariff_Class>1]~
  AIMS_Directions$threat[diffTariff_Class>1],
  yaxt="n",horizontal=TRUE,
  xlab="number of tariff classes changed",
  ylab="Threatening Direction")
axis(2,1:2,c("No", "Yes"),las=1)
dev.off()
fisher.test(table(diffTariff_ClassANY,AIMS_Directions$threat))

pdf("Line_boxplot.pdf",width=7, height=4)
boxplot(abs(diffLine_Nbr)[abs(diffLine_Nbr)>1]
  ~AIMS_Directions$threat[abs(diffLine_Nbr)>1],
  yaxt="n",horizontal=TRUE,
  xlab="number of line differences",
  ylab="Threatening Direction")
axis(2,1:2,c("No", "Yes"),las=1)
dev.off()

pdf("Goods_Quality_boxplot.pdf",width=7, height=4)
boxplot(diffGoods_Quantity_Value_1[diffGoods_Quantity_Value_1>1]~
  AIMS_Directions$threat[diffGoods_Quantity_Value_1>1],
  yaxt="n",horizontal=TRUE,
  xlab="generalized Levenshtein distance",
  ylab="Threatening Direction")
axis(2,1:2,c("No", "Yes"),las=1)
dev.off()
fisher.test(table(diffGoods_Quantity_Value_1ANY,AIMS_Directions$threat))

pdf("Goods_Description_boxplot.pdf",width=7, height=4)
boxplot(diffGoods_Description[diffGoods_Description>1]~
  AIMS_Directions$threat[diffGoods_Description>1],
  yaxt="n",horizontal=TRUE,
  xlab="generalized Levenshtein distance",
  ylab="Threatening Direction")
axis(2,1:2,c("No", "Yes"),las=1)
dev.off()
fisher.test(table(diffGoods_DescriptionANY,AIMS_Directions$threat))

#for each amendment, does it change High Risk Answer rate or incidence?
diffCPHR<-diffCPHR_prop<-afterCPHR<-beforeCPHR<-vector()
for(i in amendrows)
{
  id<-AIMS_Directions$Quarantine.Entry[i] #entry id for that direction

  #getting line details and selecting two time points
  tempi<- subset(CP, CP$Declaration_Id==id)
  beforei<-subset(tempi,tempi$FID_Transaction_Msg_Id==beforedate[i])
  afteri<-subset(tempi,tempi$FID_Transaction_Msg_Id==afterdate[i])

  #assessing if the overall incidence of a High Risk Answer changes
  beforeCPHR[i]<-1*(length(beforei$High_Risk_Answer_Type[beforei$High_Risk_Answer_Type=="Y"])>0)
  afterCPHR[i]<-1*(length(afteri$High_Risk_Answer_Type[afteri$High_Risk_Answer_Type=="Y"])>0)
  diffCPHR[i]<-afterCPHR[i]-beforeCPHR[i]

  #assessing if the rate of incidence of a High Risk Answer across lines changes

```

```

nline<-unique(beforei$Line_Nbr)
HRcountb<-HRcounta<-0
for(j in nline)
{
  #assess changes at line level
  tempb<-beforei[beforei$Line_Nbr==j,"High_Risk_Answer_Type"]
  HRcountb<-HRcountb+1*("Y" %in% tempb)
}
beforeCPHR_prop<-HRcountb/length(nline)

nline<-unique(afteri$Line_Nbr)
for(j in nline)
{
  #assess changes at line level
  tempa<-afteri[afteri$Line_Nbr==j,"High_Risk_Answer_Type"]
  HRcounta<-HRcounta+1*("Y" %in% tempa)
}
afterCPHR_prop<-HRcounta/length(nline)

diffCPHR_prop[i]<-afterCPHR_prop-beforeCPHR_prop
}

#####
#####Special AIMS Directions#####
#####

ndir<-dim(AIMS_Directions)[1]
#indicating special directions or not
threats1<-c("Inspect (Hold Seals Intact)",
            "Inspect (unpack)",
            "SIP - Hold Seals Intact",
            "SIP - Inspect (Unpack)",
            "AEP Random Audit",
            "Follow Up Inspection Required"
)
nthr<-length(threats1)

for(i in 1:ndir)
{
  AIMS_Directions$threat[i]<-1*(AIMS_Directions$Direction[i] %in% threats1)
}

#####
#amendments#
#####

#initialising
beforedate<-afterdate<-vector()
broker<-vector()
class(beforedate) <- class(afterdate) <- class(AIMS_Directions$Initiatingdate)
for(i in 1:ndir)
{
  id<-AIMS_Directions$Quarantine.Entry[i] #entry id for that direction
  dt<-AIMS_Directions$Initiatingdate[i] #date of the direction
  #selecting details for that direction
  temp<- subset(CEBRA_decs_with_line_details,

```

```

    CEBRA_decs_with_line_details$Declaration_Id==id)
broker[i]<-temp$Brokerage_Id[1]
datesi<-unique(temp$FID_Transaction_Msg_Id)
ndate<-length(datesi)
if(ndate>1)
{

    if(dt<datesi[ndate])
    {
        for(j in 1:(ndate-1))
        {
            gap1<-difftime(dt,datesi[j], unit="days")
            if((gap1>0)&(gap1<30)){beforedate[i]<-datesi[j]}
            if((gap1>0)&(gap1<30)){afterdate[i]<-datesi[j+1]}
        }
    }
}
#diff1<-dim(AIMS_Directions)[1]-length(beforedate)
#beforedate1<-c(beforedate,rep(0,diff1))
#afterdate1<-c(afterdate,rep(0,diff1))
write.csv(data.frame(AIMS_Directions$Quarantine.Entry,
    AIMS_Directions$Initiatingdate,beforedate,afterdate),"AmendDates1.csv")

#just a count of corresponding amendments
AmendAfter<-1*(afterdate>0)
AmendAfter[is.na(AmendAfter)] <- 0

#testing if special Directions result in more amendments:
fisher.test(table(AmendAfter,AIMS_Directions$threat))

#overall patterns of directions
tab1<-prop.table(table(AIMS_Directions$Direction,AmendAfter),1)
tab1<-cbind(tab1,table(AIMS_Directions$Direction))
tab1[order(tab1[,2],decreasing=TRUE),]
write.csv(tab1[order(tab1[,2],decreasing=TRUE),],"AmendbyDirection1.csv")

#testing time gap
#gap1<-afterdate-AIMS_Directions$Initiatingdate

##breaking down by type of amendment
#STILL TO ADD:
#quantity
#high risk answer to CP

#just relevant rows
amendrows<-which(AmendAfter==1)

#look at the before and after date for each direction and which
diffDelivery_Address<-diffImporter_Id<-diffImporter_Address<-vector()
changeSupplier_Id<-changeImporter_Id<-changeBrokerage_Id<-diffTariff_Class<-vector()
diffGoods_Quantity_Value_1<-diffGoods_Description<-vector()
diffLine_Nbr<-addLine_Nbr<-deleteLine_Nbr<-vector()
for(i in amendrows)
{
    id<-AIMS_Directions$Quarantine.Entry[i] #entry id for that direction

```



```

#getting line details and selecting two time points
tempi<- subset(CEBRA_decs_with_line_details, CEBRA_decs_with_line_details$Declaration_Id==id)
beforei<-subset(tempi,tempi$FID_Transaction_Msg_Id==beforedate[i])
afteri<-subset(tempi,tempi$FID_Transaction_Msg_Id==afterdate[i])

diffLine_Nbr[i]<-length(unique(afteri$Line_Nbr))-length(unique(beforei$Line_Nbr))
addLine_Nbr[i]<-length(setdiff(unique(afteri$Line_Nbr),
  unique(beforei$Line_Nbr)))
deleteLine_Nbr[i]<-length(setdiff(unique(beforei$Line_Nbr),
  unique(afteri$Line_Nbr)))
diffTariff_Class[i]<-length(setdiff(unique(beforei$Tariff_Class),
  unique(afteri$Tariff_Class)))
diffGoods_Description[i]<-length(setdiff(unique(beforei$Goods_Description),
  unique(afteri$Goods_Description)))
diffGoods_Quantity_Value_1[i]<-length(setdiff(unique(beforei$Quantity_Value_1),
  unique(afteri$Quantity_Value_1)))

#focusing on line 1 only for entry level characteristics
beforei1<-beforei[1,]
afteri1<-afteri[1,]

diffDelivery_Address[i]<-adist(afteri1$Delivery_Address,beforei1$Delivery_Address)
diffImporter_Address[i]<-adist(afteri1$Importer_Address,beforei1$Importer_Address)
changeImporter_Id[i]<-1-(afteri1$Importer_Id==beforei1$Importer_Id)
changeSupplier_Id[i]<-1-(afteri1$Supplier_Id==beforei1$Supplier_Id)
changeBrokerage_Id[i]<-1-(afteri1$Brokerage_Id==beforei1$Brokerage_Id)

}

addLineANY_Nbr<-1*(addLine_Nbr>0)
deleteLineANY_Nbr<-1*(deleteLine_Nbr>0)
diffLineANYPOS_Nbr<-1*(diffLine_Nbr>0)
diffLineANYNEG_Nbr<-1*(diffLine_Nbr<0)
diffDelivery_AddressANY<-1*(diffDelivery_Address>0)
diffImporter_AddressANY<-1*(diffImporter_Address>0)
diffTariff_ClassANY<-1*(diffTariff_Class>0)
diffGoods_Quantity_Value_1ANY<-1*(diffGoods_Quantity_Value_1>0)
diffGoods_DescriptionANY<-1*(diffGoods_Description>0)

fisher.test(table(changeImporter_Id,AIMS_Directions$threat))
fisher.test(table(changeSupplier_Id,AIMS_Directions$threat))
fisher.test(table(changeBrokerage_Id,AIMS_Directions$threat))
fisher.test(table(addLineANY_Nbr,AIMS_Directions$threat))
fisher.test(table(deleteLineANY_Nbr,AIMS_Directions$threat))
fisher.test(table(diffLineANYPOS_Nbr,AIMS_Directions$threat))
fisher.test(table(diffLineANYNEG_Nbr,AIMS_Directions$threat))
pdf("Delivery_boxplot1.pdf",width=7, height=4)
boxplot(diffDelivery_Address[diffDelivery_Address>1]~
  AIMS_Directions$threat[diffDelivery_Address>1],
  yaxt="n",horizontal=TRUE,
  xlab="generalized Levenshtein distance",
  ylab="Special Direction")
axis(2,1:2,c("No","Yes"),las=1)
dev.off()
fisher.test(table(diffDelivery_AddressANY,AIMS_Directions$threat))

pdf("Importer_boxplot1.pdf",width=7, height=4)
boxplot(diffImporter_Address[diffImporter_Address>1]~

```

```

    AIMS_Directions$threat[diffImporter_Address>1],
    yaxt="n",horizontal=TRUE,
    xlab="generalized Levenshtein distance",
    ylab="Special Direction")
axis(2,1:2,c("No", "Yes"),las=1)
dev.off()
fisher.test(table(diffImporter_AddressANY,AIMS_Directions$threat))

pdf("Tariff_boxplot1.pdf",width=7, height=4)
boxplot(diffTariff_Class[diffTariff_Class>1]~
    AIMS_Directions$threat[diffTariff_Class>1],
    yaxt="n",horizontal=TRUE,
    xlab="number of tariff classes changed",
    ylab="Special Direction")
axis(2,1:2,c("No", "Yes"),las=1)
dev.off()
fisher.test(table(diffTariff_ClassANY,AIMS_Directions$threat))

pdf("Line_boxplot1.pdf",width=7, height=4)
boxplot(abs(diffLine_Nbr)[abs(diffLine_Nbr)>1]~
    AIMS_Directions$threat[abs(diffLine_Nbr)>1],
    yaxt="n",horizontal=TRUE,
    xlab="number of line differences",
    ylab="Special Direction")
axis(2,1:2,c("No", "Yes"),las=1)
dev.off()

pdf("Goods_Quality_boxplot1.pdf",width=7, height=4)
boxplot(diffGoods_Quantity_Value_1[diffGoods_Quantity_Value_1>1]~
    AIMS_Directions$threat[diffGoods_Quantity_Value_1>1],
    yaxt="n",horizontal=TRUE,
    xlab="generalized Levenshtein distance",
    ylab="Special Direction")
axis(2,1:2,c("No", "Yes"),las=1)
dev.off()
fisher.test(table(diffGoods_Quantity_Value_1ANY,AIMS_Directions$threat))

pdf("Goods_Description_boxplot1.pdf",width=7, height=4)
boxplot(diffGoods_Description[diffGoods_Description>1]~
    AIMS_Directions$threat[diffGoods_Description>1],
    yaxt="n",horizontal=TRUE,
    xlab="generalized Levenshtein distance",
    ylab="Special Direction")
axis(2,1:2,c("No", "Yes"),las=1)
dev.off()
fisher.test(table(diffGoods_DescriptionANY,AIMS_Directions$threat))

#for each amendment, does it change High Risk Answer rate or incidence?
diffCPHR<-diffCPHR_prop<-afterCPHR<-beforeCPHR<-vector()
for(i in amendrows)
{
    id<-AIMS_Directions$Quarantine.Entry[i] #entry id for that direction

    #getting line details and selecting two time points
    tempi<- subset(CP, CP$Declaration_Id==id)
    beforei<-subset(tempi,tempi$FID_Transaction_Msg_Id==beforedate[i])
    afteri<-subset(tempi,tempi$FID_Transaction_Msg_Id==afterdate[i])

```

```

#assessing if the overall incidence of a High Risk Answer changes
beforeCPHR[i]<-1*(length(beforei$High_Risk_Answer_Type[beforei$High_Risk_Answer_Type=="Y"])>0)
afterCPHR[i]<-1*(length(afteri$High_Risk_Answer_Type[afteri$High_Risk_Answer_Type=="Y"])>0)
diffCPHR[i]<-afterCPHR[i]-beforeCPHR[i]

#assessing if the rate of incidence of a High Risk Answer across lines changes
nline<-unique(beforei$Line_Nbr)
HRcountb<-HRcounta<-0
for(j in nline)
{
  #assess changes at line level
  tempb<-beforei[beforei$Line_Nbr==j,"High_Risk_Answer_Type"]
  HRcountb<-HRcountb+1*("Y" %in% tempb)
}
beforeCPHR_prop<-HRcountb/length(nline)

nline<-unique(afteri$Line_Nbr)
for(j in nline)
{
  #assess changes at line level
  tempa<-afteri[afteri$Line_Nbr==j,"High_Risk_Answer_Type"]
  HRcounta<-HRcounta+1*("Y" %in% tempa)
}
afterCPHR_prop<-HRcounta/length(nline)

diffCPHR_prop[i]<-afterCPHR_prop-beforeCPHR_prop
}

#exploring the characteristics of amendments that remove high risk
CPHRremoval<-1*(diffCPHR==1)

fisher.test(table(changeImporter_Id,CPHRremoval))
fisher.test(table(changeSupplier_Id,CPHRremoval))
fisher.test(table(addLineANY_Nbr,CPHRremoval))
fisher.test(table(deleteLineANY_Nbr,CPHRremoval))
fisher.test(table(diffLineANYPOS_Nbr,CPHRremoval))
fisher.test(table(diffLineANYNEG_Nbr,CPHRremoval))
fisher.test(table(diffDelivery_AddressANY,CPHRremoval))
fisher.test(table(diffImporter_AddressANY,CPHRremoval))
fisher.test(table(diffTariff_ClassANY,CPHRremoval))
fisher.test(table(diffGoods_DescriptionANY,CPHRremoval))
fisher.test(table(diffGoods_Quantity_Value_1ANY,CPHRremoval))

table(changeImporter_Id,CPHRremoval)
table(changeSupplier_Id,CPHRremoval)
table(addLineANY_Nbr,CPHRremoval)
table(deleteLineANY_Nbr,CPHRremoval)
table(diffLineANYPOS_Nbr,CPHRremoval)
table(diffLineANYNEG_Nbr,CPHRremoval)
table(diffDelivery_AddressANY,CPHRremoval)
table(diffImporter_AddressANY,CPHRremoval)
table(diffTariff_ClassANY,CPHRremoval)
table(diffGoods_DescriptionANY,CPHRremoval)
table(diffGoods_Quantity_Value_1ANY,CPHRremoval)

```

## Appendix E

# Risk Factor Extraction with VMS

### E.1 VMSscript.R

```
require(car)
require(tree)
require(rpart)
require(ROCR)
require(gbm)
require(randomForest)
require(stringr)

insp<-read.csv("data/inspections.csv")

#####
#### initial data cleaning ####
#####

insp1<-insp[insp$inspect.type=="ROUTINE",]
insp1<-insp1[insp1$inspect.result!="NO RESULT",]
insp1<-insp1[insp1$inspect.result!="NON-INSPECT",]
insp1$inspect.result<-factor(insp1$inspect.result)
insp1$inspect.result1<-car::recode(insp1$inspect.result,
  "'FAIL'=1;'NON-CONFORMITY'=1;'PASS'=0")

#Adding agents
voyage1<-read.csv("voyage_results_org_for_CEBRA.csv")
voyage1<-voyage1[,c("agent.name", "inspect.id",
  "imo.number", "inspection.date.time")]
colnames(voyage1)[4]<-"inspect.date.time"

#sorting out differing date codes
dt1<-voyage1$inspect.date.time
dt1<-as.POSIXct(strptime(dt1,"%d/%m/%Y %H:%M"))
voyage1$inspect.date<-format(dt1,format="%d/%m/%Y")
voyage1$inspect.time<-format(dt1,format="%H:%M")
dt2<-insp1$inspect.date.time
dt2<-as.POSIXct(strptime(dt2, "%Y-%m-%d %H:%M:%S"))
insp1$inspect.date<-format(dt2,format="%d/%m/%Y")
insp1$inspect.time<-format(dt2,format="%H:%M")

#matching to inspections
insp1<-merge(insp1,voyage1,by=c("inspect.date",
  "inspect.time", "imo.number"), all.x=T)
rm(voyage1, dt1, dt2, insp)
```

```

#dealing with the unmatched
insp1$agent.name<-factor(insp1$agent.name,
  levels=c(levels(insp1$agent.name), "no.match"))
insp1$agent.name[is.na(insp1$agent.name)]<-"no.match"

#preparing variables
insp1$agent.name<-factor(insp1$agent.name)
insp1$vessel.type<-factor(insp1$vessel.type)
insp1$inspect.month<-factor(insp1$inspect.month)
insp1$inspect.quarter<-factor(insp1$inspect.quarter)
insp1$inspect.year<-factor(insp1$inspect.year)
insp1$proclaimedport<-factor(insp1$proclaimedport)
insp1$current.port<-factor(insp1$current.port)
insp1$regioncode<-factor(insp1$regioncode)
insp1$vessel.name<-factor(insp1$vessel.name)
insp1$pratique.visit<-factor(insp1$pratique.visit)
insp1$pdccycle<-factor(insp1$pdccycle, exclude = NULL)
#insp1$pdccchangereason<-factor(insp1$pdccchangereason, exclude = NULL)
#insp1$biofoulingproblem<-factor(insp1$biofoulingproblem) #only one level
#insp1$faildesc<-factor(insp1$faildesc, exclude = NULL) #feature of failure
#insp1$pdcc<-factor(insp1$pdcc) #only one level
insp1$visit.is.last<-factor(insp1$visit.is.last)
insp1$visit.is.first<-factor(insp1$visit.is.first)
insp1$visit.month<-factor(insp1$visit.month)
insp1$visit.quarter<-factor(insp1$visit.quarter)
insp1$visit.year<-factor(insp1$visit.year)
insp1$visit.ship.type<-factor(insp1$visit.ship.type)
insp1$visit.length.category<-factor(insp1$visit.length.category)
insp1$visit.visit.vessel.class<-factor(insp1$visit.vessel.class)
insp1$pdcc.voyage<-factor(insp1$pdcc.voyage)
insp1$pdccvisitcount<-factor(insp1$pdccvisitcount)

#functions to collapse and create dummy variables
source('vms_functions.r')

#group with less than 20 counts, 20-99 counts and 100-199
insp1$last.port1<-factor(collapsing.variables3(insp1$last.port,20,100,200))
insp1$vessel.name1<-factor(collapsing.variables3(insp1$vessel.name,20,100,200))

#group with less than 20 counts, 20-99 counts
insp1$last.country1<-factor(collapsing.variables2(insp1$last.country,20,100))
insp1$current.port1<-factor(collapsing.variables2(insp1$current.port,20,100))
insp1$agent.name1<-factor(collapsing.variables2(insp1$agent.name,20,100))

#creating dummy variables
agent.name2<-dummifying.variables(insp1$agent.name1,"agent.name")
vessel.type2<-dummifying.variables(insp1$vessel.type,"vessel.type")
inspect.month2<-dummifying.variables(insp1$inspect.month,"inspect.month")
inspect.quarter2<-dummifying.variables(insp1$inspect.quarter,"inspect.quarter")
inspect.year2<-dummifying.variables(insp1$inspect.year,"inspect.year")
proclaimedport2<-dummifying.variables(insp1$proclaimedport,"proclaimedport")
current.port2<-dummifying.variables(insp1$current.port1,"current.port")
regioncode2<-dummifying.variables(insp1$regioncode,"regioncode")
vessel.name2<-dummifying.variables(insp1$vessel.name1,"vessel.name")
pratique.visit2<-dummifying.variables(insp1$pratique.visit,"pratique.visit")
pdccycle2<-dummifying.variables(insp1$pdccycle,"pdccycle")
visit.is.last2<-dummifying.variables(insp1$visit.is.last,"visit.is.last")

```

```

visit.is.first2<-dummmifying.variables(insp1$visit.is.first,"visit.is.first")
visit.month2<-dummmifying.variables(insp1$visit.month,"visit.month")
visit.quarter2<-dummmifying.variables(insp1$visit.quarter,"visit.quarter")
visit.year2<-dummmifying.variables(insp1$visit.year,"visit.year")
visit.ship.type2<-dummmifying.variables(insp1$visit.ship.type,"visit.ship.type")
visit.length.category2<-dummmifying.variables(insp1$visit.length.category,
"visit.length.category")
visit.vessel.class2<-dummmifying.variables(insp1$visit.vessel.class,
"visit.vessel.class")
pdc.voyage2<-dummmifying.variables(insp1$pdc.voyage,"pdc.voyage")
last.country2<-dummmifying.variables(insp1$last.country1,"last.country")
last.port2<-dummmifying.variables(insp1$last.port1,"last.port")

insp2<-data.frame(insp1,agent.name2,vessel.type2, inspect.month2,
inspect.quarter2, inspect.year2, proclaimedport2,
current.port2, regioncode2, vessel.name2, pratique.visit2,
pdccycle2, visit.is.last2, visit.is.first2,
visit.month2, visit.quarter2, visit.year2, visit.ship.type2,
visit.length.category2, visit.vessel.class2, pdc.voyage2,
last.country2, last.port2)

vars<-c(colnames(agent.name2),colnames(vessel.type2),
colnames(inspect.month2),colnames(inspect.quarter2),
colnames(inspect.year2),colnames(proclaimedport2),
colnames(current.port2),colnames(regioncode2),
colnames(vessel.name2),colnames(pratique.visit2),
colnames(pdccycle2),colnames(visit.is.last2),
colnames(visit.is.first2),colnames(visit.month2),
colnames(visit.quarter2),colnames(visit.year2),
colnames(visit.ship.type2),colnames(visit.length.category2),
colnames(visit.vessel.class2),colnames(pdc.voyage2),
colnames(last.country2),colnames(last.port2))

fvars<-as.formula(paste("inspect.result1~",paste(vars,collapse="+")))

save(list=c("insp2","vars"),file="vms_insp_for_analysis.RData")
rm(insp1)

#####
#### initial variable selection ####
#####

#using AUC

# split into test and train
N <- dim(insp2)[1]
idx <- (runif(N) < .5)
insp2.train <- insp2[idx==T,]
insp2.test <- insp2[idx==F,]

AUCvars<-vector()
for(i in 1:length(vars))
{
  vari<-vars[i]
  fvari<-as.formula(paste("inspect.result1~",vari))
  glmi<-glm(fvari, family = binomial("logit"),data=insp2.train)
  predi<-predict(glmi,newdata=insp2.test)
  predictioni<-prediction(predi,(insp2.test$inspect.result1==1))

```

```

    temp1<-round(attributes(performance(predictioni, 'auc'))$y.values[[1]],4)
    AUCvars[i]<-temp1
  }
plot(sort(AUCvars,decreasing=T))
rm(insp2.train,insp2.test)

#new vars

vars2<-vars[AUCvars>0.505]
fvars2<-as.formula(paste("inspect.result1~",paste(vars2,collapse="+")))

#####
#### GBM modelling ####
#####

gbmfit1 <- gbm(fvars2,distribution="bernoulli",data=insp2,n.trees=2000,
  interaction.depth=4, shrinkage=0.01, cv.folds = 5)
ntrees <- gbm.perf(gbmfit2,method="cv")
table1<-summary(gbmfit1, n.trees=ntrees)

vars3<-as.character(table1[,1])
n3<-length(vars3)
n4<-dim(insp2)[1]
data4<-insp2[,vars3]
rate<-rate2<-rarity<-odds<-rep(0,n3)
for(i in 1:n3)
{
totalyi<-table(insp2$inspect.result1,data4[,i])[1,2]+
  table(insp2$inspect.result1,data4[,i])[2,2]
totalnotyi<-table(insp2$inspect.result1,data4[,i])[2,1]+
  table(insp2$inspect.result1,data4[,i])[1,1]
totalnci<-table(insp2$inspect.result1,data4[,i])[2,2]
totalncnoti<-table(insp2$inspect.result1,data4[,i])[2,1]
rate[i]<-round(totalnci/totalyi,digits=3)
rate2[i]<-round(totalncnoti/totalnotyi,digits=3)
rarity[i]<-table(data4[,i])[2]/n4
odds[i]<-rate[i]*(1-rate2[i])/(1-rate[i])/rate2[i]
}
table2<-data.frame(Variable=table1[,1],
  Importance_GBM=round(table1[,2],digits=3),
  Rate_NonComp_T=rate,Rate_NonComp_F=rate2, Odds<-round(odds,digits=2),
  Rarity=round(100*rarity, digits=1))

```

## E.2 vmsfunctions.R

```
#####
#### general utils ####
#####

collapsing.variables3<-function(var1,a,b,c){
  vartab1<-table(var1)
  levels1<-names(vartab1)
  freqs<-as.numeric(vartab1)
  var1group<-vector()
  for(j in 1:length(levels1))
  {
    if(freqs[j]<a){var1group[j]<-"lowest"}
    else
    {
      if(freqs[j]<b){var1group[j]<-"lower"}
      else
      {
        if(freqs[j]<c){var1group[j]<-"low"}
        else
        {
          var1group[j]<-levels1[j]
        }
      }
    }
  }
  var1codes<-data.frame(levels1,var1group)

  group<-vector()
  for(i in 1:length(var1))
  {
    vartemp<-var1[i]
    group[i]<-as.character(var1codes$var1group[levels1==vartemp])
  }

  return(group)
}

collapsing.variables2<-function(var1,a,b){
  vartab1<-table(var1)
  levels1<-names(vartab1)
  freqs<-as.numeric(vartab1)
  var1group<-vector()
  for(j in 1:length(levels1))
  {
    if(freqs[j]<a){var1group[j]<-"lowest"}
    else
    {
      if(freqs[j]<b){var1group[j]<-"lower"}
      else
      {
        var1group[j]<-levels1[j]
      }
    }
  }
  group<-vector()
  for(i in 1:length(var1))
```



```

{
  vartemp<-var1[i]
  if(is.na(vartemp)){group[i]<-NA}
  if(is.na(vartemp)==F){
    group[i]<-as.character(var1group[levels1==vartemp])}
  }
  return(group)
}

dummifying.variables<-function(x,y){
  temp<- model.matrix(~., data = data.frame(x))
  colnames(temp)[1]<-names(table(x))[1]
  colnames(temp) <- gsub("[:punct:]", "", colnames(temp))
  colnames(temp) <- gsub(" ", ".", colnames(temp))
  colnames(temp) <- sub("x","", colnames(temp))
  colnames(temp) <- paste(y,colnames(temp),sep="")
  return(temp)
}

```

## E.3 Performance Indicators for Cargo Compliance Verification

### E.3.1 Tables of results for July 2013

Tables E.1 through to E.8 provide the CCV results by processing state, country, tariff, broker, importer and supplier. For simplicity, those levels with no errors were excluded in each table, and all zero cells are left blank. The actual names of the brokers, importers and suppliers have been suppressed for simplicity, as these results are illustrative only. This analysis was restricted to line entries only, and all food profiles were excluded.

As each entry has multiple lines, it is possible for there to be entries that have lines with different profile pathways. Hence the number of entries with each pathway doesn't add to the total number of entries. Failure rates are by line, not by entry.

**Table E.1:** CCV failure rates by profile pathway

	Overall	Commodity	Non-Com	Both	Lines	Entries
Overall	0.026	0.005	0.012	0.010	7594	705
high risk	0.069	0.027	0.018	0.024	881	391
sometimes referred	0.055	0.006	0.028	0.020	495	131
downgraded profile	0.023	0.001	0.011	0.011	3990	260
not profiled	0.009	0.002	0.007		2228	182

**Table E.2:** CCV failure rates by country

	Overall	Commodity	Non-Com	Both	Lines	Entries
IRAN	0.833	0.833			6	2
INDIA	0.127	0.054	0.074		204	37
MALAYSIA	0.097	0.014	0.083		72	28
AUSTRIA	0.091		0.091		22	8
JAPAN	0.056			0.056	719	31
NEW ZEALAND	0.052		0.052		248	55
CHINA	0.051	0.009	0.023	0.018	1859	294
CHILE	0.023		0.023		43	15
FRANCE	0.015		0.015		68	19
TAIWAN	0.012	0.004	0.008		256	40
INDONESIA	0.011		0.011		186	43
UNITED STATES	0.003		0.003		1218	59
THAILAND	0.003			0.003	307	46

**Table E.3:** CCV failure rate by processing state

	Overall	Commodity	Non-Com	Both	Lines	Entries
NSW	0.030	0.000	0.004	0.026	2905	238
NT					1	1
QLD	0.098	0.032	0.066		870	139
SA	0.020	0.017	0.003		297	35
TAS					95	15
VIC	0.002	0.000	0.002		2429	213
WA	0.015		0.015		997	64

**Table E.4:** CCV failure rate by broker

	Overall	Commodity	Non-Com	Both	Lines	Entries
A	1.000		1.000		13	1
B	1.000		1.000		7	1
C	1.000		1.000		1	1
D	1.000	1.000			1	1
E	1.000		1.000		1	1
F	1.000		1.000		1	1
G	0.833			0.833	54	2
H	0.811			0.811	37	5
I	0.600		0.600		25	4
J	0.500		0.500		2	2
K	0.500		0.500		2	2
L	0.500		0.500		2	2
M	0.423	0.423			26	7
N	0.333		0.333		3	3
O	0.267	0.267			15	4
P	0.235		0.235		34	4
Q	0.167	0.167			6	4
R	0.154		0.154		39	8
S	0.143		0.143		7	3
T	0.132		0.132		38	10
U	0.118	0.059	0.059		17	6
V	0.111	0.111			45	9
W	0.095	0.074	0.021		95	13
X	0.083		0.083		12	4
Y	0.077	0.077			13	7
Z	0.072	0.008	0.064		125	26
AA	0.063		0.063		16	5
AB	0.043		0.043		23	8
AC	0.040		0.040		25	4
AD	0.038		0.038		52	7
AE	0.038		0.038		26	9
AF	0.017	0.017			58	4
AG	0.011		0.011		92	6
AH	0.010	0.010			191	28
AI	0.010		0.010		103	4
AJ	0.005		0.005		365	18
AK	0.003		0.003		927	40
AL	0.002		0.002		1021	22

**Table E.5:** CCV failure rate by importer code

	Overall	Commodity	Non-Com	Both	Lines	Entries
A	1.000			1.000	30	1
B	1.000		1.000		13	1
C	1.000		1.000		8	1
D	1.000		1.000		8	1
E	1.000		1.000		7	1
F	1.000	1.000			7	1
G	1.000	1.000			5	1
H	1.000		1.000		5	1
I	1.000		1.000		5	1
J	1.000		1.000		2	1
K	1.000		1.000		2	1
L	1.000		1.000		2	1
M	1.000		1.000		1	1
N	1.000		1.000		1	1
O	1.000		1.000		1	1
P	1.000		1.000		1	1
Q	1.000		1.000		1	1
R	1.000		1.000		1	1
S	1.000	1.000			1	1
T	1.000	1.000			1	1
U	1.000		1.000		1	1
V	1.000		1.000		1	1
W	1.000		1.000		1	1
X	1.000	1.000			1	1
Y	1.000		1.000		1	1
Z	1.000		1.000		1	1
AA	1.000		1.000		1	1
AB	0.882		0.882		17	2
AC	0.833			0.833	54	2
AD	0.579	0.579			19	2
AE	0.571	0.571			7	2
AF	0.500		0.500		2	2
AG	0.500		0.500		2	2
AH	0.500		0.500		2	2
AI	0.500		0.500		2	2
AJ	0.400	0.400			5	2
AK	0.333		0.333		3	3
AL	0.333	0.333			3	1
AM	0.250	0.250			4	1
AN	0.154		0.154		13	4
AO	0.079		0.079		38	12
AP	0.021	0.021			47	1

**Table E.6:** CCV failure rate by profile code

	Overall	Commodity	Non-Com	Both	Lines	Entries
A	1.000	1.000		1.000	2	1
B	1.000			1.000	2	1
C	1.000				1	1
D	1.000			1.000	1	1
E	1.000			1.000	1	1
F	1.000		1.000		1	1
G	0.742			0.742	31	8
H	0.714		0.643	0.071	14	4
I	0.600			0.600	5	3
J	0.545		0.500	0.045	22	5
K	0.500		0.500		2	2
L	0.474		0.053	0.421	19	9
M	0.435		0.043	0.391	23	12
N	0.400			0.400	15	9
O	0.400			0.400	10	5
P	0.333		0.222	0.111	9	6
Q	0.333		0.222	0.111	9	6
R	0.333		0.222	0.111	9	6
S	0.333		0.333		6	4
T	0.333		0.333		6	4
U	0.333		0.333		6	4
V	0.333			0.333	3	3
W	0.263		0.263		19	9
X	0.250		0.250		8	5
Y	0.250			0.250	4	4
Z	0.250			0.250	4	2
AA	0.250		0.250		4	4
AB	0.250			0.250	4	3
AC	0.250			0.250	4	4
AD	0.250			0.250	4	4
AE	0.250			0.250	4	2
AF	0.250			0.250	4	4
AG	0.229			0.229	48	22
AH	0.222		0.222		9	6
AI	0.222		0.222		9	6
AJ	0.222		0.222		9	6
AK	0.207		0.034	0.172	29	10
AL	0.200		0.200		5	4

## CCV failure rate by profile code (continued...)

	Overall	Commodity	Non-Com	Both	Lines	Entries
AM	0.200			0.200	5	2
AN	0.200		0.200		5	3
AO	0.200		0.200		5	4
AP	0.200		0.200		5	4
AQ	0.173		0.135	0.038	52	18
AR	0.160	0.080		0.080	25	12
AS	0.143		0.143		7	5
AT	0.143		0.143		7	6
AU	0.130	0.065	0.022	0.043	46	20
AV	0.128		0.021	0.106	47	12
AW	0.125			0.125	8	4
AX	0.118	0.059	0.059		17	10
AY	0.111		0.111		9	5
AZ	0.111		0.111		9	6
BA	0.100			0.100	10	5
BB	0.097		0.097		31	11
BC	0.097		0.097		31	11
BD	0.084	0.066	0.018		332	107
BE	0.083			0.083	36	8
BF	0.083			0.083	36	8
BG	0.083			0.083	36	8
BH	0.083		0.083		12	9
BI	0.076	0.015	0.061		66	11
BJ	0.065	0.009	0.037	0.019	107	23
BK	0.064		0.034	0.030	203	47
BL	0.061		0.061		114	18
BM	0.061		0.061		114	18
BN	0.055			0.055	55	15
BO	0.055			0.055	55	15
BP	0.055			0.055	55	15
BQ	0.055			0.055	55	15
BR	0.055			0.055	55	15
BS	0.049			0.049	41	12
BT	0.049			0.049	41	12
BU	0.048			0.048	21	13
BV	0.036		0.036		28	16
BW	0.033		0.033		60	31
BX	0.032			0.032	93	17

## CCV failure rate by profile code (continued...)

	Overall	Commodity	Non-Com	Both	Lines	Entries
BY	0.032		0.032		31	24
BZ	0.032		0.032		31	24
CA	0.030		0.030		33	7
CB	0.030		0.030		33	7
CC	0.030		0.030		67	24
CD	0.027		0.027		37	17
CE	0.022		0.022		46	19
CF	0.021		0.021		48	32
CG	0.020		0.020		51	39
CH	0.019		0.019		52	21
CI	0.013		0.013		149	50
CJ	0.013		0.013		149	50
CK	0.013		0.013		80	22
CL	0.012		0.012		85	26
CM	0.008	0.003	0.005		778	193
CN	0.005	0.003	0.002		658	123
CO	0.005	0.003	0.002		660	117
CP	0.005	0.003	0.002		660	117
CQ	0.005	0.003	0.002		660	117
CR	0.003	0.003			310	27
CS	0.003	0.003			351	46

**Table E.7:** CCV failure rate by supplier code

	Overall	Commodity	Non-Com	Both	Lines	Entries
A	1.000			1.000	45	1
B	1.000			1.000	30	1
C	1.000		1.000		15	1
D	1.000		1.000		8	1
E	1.000		1.000		8	1
F	1.000		1.000		7	1
G	1.000		1.000		5	1
H	1.000		1.000		5	1
I	1.000		1.000		4	1
J	1.000	1.000			4	1
K	1.000	1.000			4	1
L	1.000		1.000		3	1
M	1.000		1.000		2	1
N	1.000		1.000		2	1
O	1.000	1.000			2	1
P	1.000		1.000		2	1
Q	1.000		1.000		1	1
R	1.000		1.000		1	1
S	1.000		1.000		1	1
T	1.000		1.000		1	1
U	1.000		1.000		1	1
V	1.000		1.000		1	1
W	1.000		1.000		1	1
X	1.000		1.000		1	1
Y	1.000		1.000		1	1
Z	1.000		1.000		1	1
AA	1.000	1.000			1	1
AB	1.000		1.000		1	1
AC	1.000		1.000		1	1
AD	1.000		1.000		1	1
AE	1.000	1.000			1	1
AF	1.000		1.000		1	1
AG	1.000	1.000			1	1
AH	1.000	1.000			1	1
AI	1.000		1.000		1	1
AJ	1.000		1.000		1	1
AK	1.000		1.000		1	1
AL	1.000		1.000		1	1
AM	0.412	0.412			17	4
AN	0.333		0.333		9	3
AO	0.333	0.333			3	1
AP	0.314	0.314			35	3
AQ	0.250		0.250		4	2
AR	0.250	0.250			4	1
AS	0.154		0.154		13	4
AT	0.062		0.062		65	2
AU	0.021	0.021			47	1



**Table E.8:** CCV failure rate by tariff code

	Overall	Commodity	Non-Com	Both	Lines	Entries
12019000	1.000			1.000	8	1
3039000	1.000			1.000	2	1
48115990	1.000		1.000		2	1
3055900	1.000			1.000	1	1
3077900	1.000			1.000	1	1
3079900	1.000			1.000	1	1
3083000	1.000			1.000	1	1
7102900	1.000		1.000		1	1
7142010	1.000			1.000	1	1
12024200	1.000		1.000		1	1
12122110	1.000			1.000	1	1
13023100	1.000			1.000	1	1
15179000	1.000			1.000	1	1
17029090	1.000		1.000		1	1
39206300	1.000		1.000		1	1
40116900	1.000	1.000			1	1
61043300	1.000		1.000		1	1
61123100	1.000		1.000		1	1
61124100	1.000		1.000		1	1
61169300	1.000		1.000		1	1
69101000	1.000		1.000		1	1
73084000	1.000		1.000		1	1
82071300	1.000		1.000		1	1
84243010	1.000		1.000		1	1
84304900	1.000		1.000		1	1
84322100	1.000		1.000		1	1
84322900	1.000		1.000		1	1
84328000	1.000		1.000		1	1
84379000	1.000	1.000			1	1
95030050	1.000			1.000	1	1
96190041	1.000		1.000		1	1
67029000	0.885			0.885	26	4
7108000	0.769		0.692	0.077	13	3
19019000	0.667	0.333	0.333		3	3
20059900	0.667	0.333		0.333	3	3
85444920	0.667		0.667		3	2
20029000	0.500	0.167	0.333		6	4
3072900	0.500			0.500	4	3
20052000	0.500		0.500		4	2

## CCV failure rate by tariff code (continued...)

	Overall	Commodity	Non-Com	Both	Lines	Entries
3038900	0.500			0.500	2	2
7103000	0.500		0.500		2	2
16051000	0.500			0.500	2	2
16055400	0.500			0.500	2	2
21021000	0.500		0.500		2	2
40116200	0.500	0.500			2	2
61052000	0.500		0.500		2	2
82071900	0.500		0.500		2	2
84283300	0.500		0.500		2	2
16055500	0.400			0.400	5	2
19049000	0.400	0.200	0.200		5	4
84314300	0.400		0.400		5	3
40119200	0.375	0.375			8	7
40129000	0.333	0.333			3	3
40139000	0.333	0.333			3	3
44079999	0.333		0.333		3	3
84289000	0.333		0.333		3	3
96151100	0.333			0.333	3	2
17049000	0.286		0.286		14	7
40119300	0.286	0.286			7	6
95059000	0.286			0.286	7	4
40116100	0.273	0.273			11	4
20049000	0.250			0.250	4	2
21041000	0.250		0.250		4	4
69089000	0.250		0.250		4	3
87087030	0.250	0.250			4	4
95069100	0.219		0.219		32	8
21039000	0.207		0.034	0.172	29	10
18069000	0.200		0.200		5	4
40119900	0.200	0.200			5	5
44123200	0.200		0.200		5	5
44182000	0.167		0.167		6	6
73089000	0.167		0.167		6	5
84332000	0.167		0.167		6	3
16042000	0.143			0.143	14	3
4090000	0.143		0.143		7	6

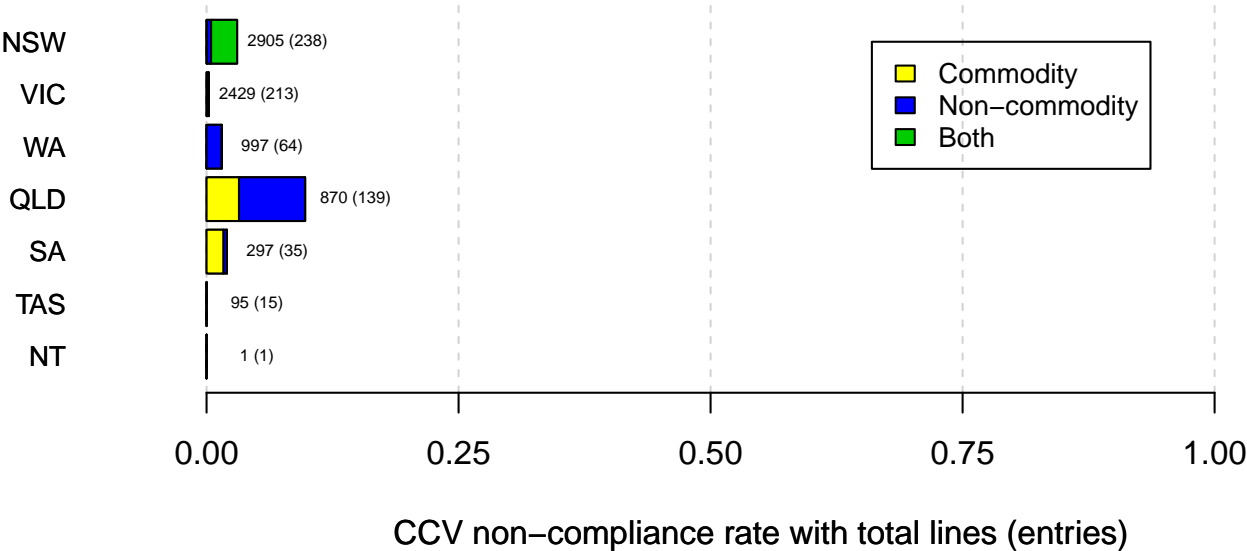
CCV failure rate by tariff code (continued...)

	Overall	Commodity	Non-Com	Both	Lines	Entries
20079900	0.143		0.143		7	6
22021000	0.143		0.143		7	4
38249090	0.143		0.143		7	6
40169100	0.143		0.143		7	6
65050090	0.143		0.143		7	5
19021900	0.125	0.042		0.083	24	12
3074900	0.125			0.125	8	7
39219090	0.111		0.111		9	8
95049090	0.111			0.111	9	4
40112000	0.085	0.069	0.015		130	62
21069090	0.075	0.019		0.057	53	15
19059000	0.066	0.009	0.038	0.019	106	23
3048900	0.063			0.063	16	9
95030070	0.061			0.061	33	5
16041900	0.056			0.056	18	4
84314990	0.043		0.043		23	8
94032000	0.042	0.042			24	16
44092900	0.040		0.040		25	24
40111000	0.038	0.013	0.025		160	69
73239900	0.036		0.036		28	15
83025000	0.029		0.029		35	21
39249000	0.027		0.027		37	17
44071010	0.023		0.023		44	18
40169900	0.021		0.021		48	32
94036000	0.015	0.008	0.008		132	35

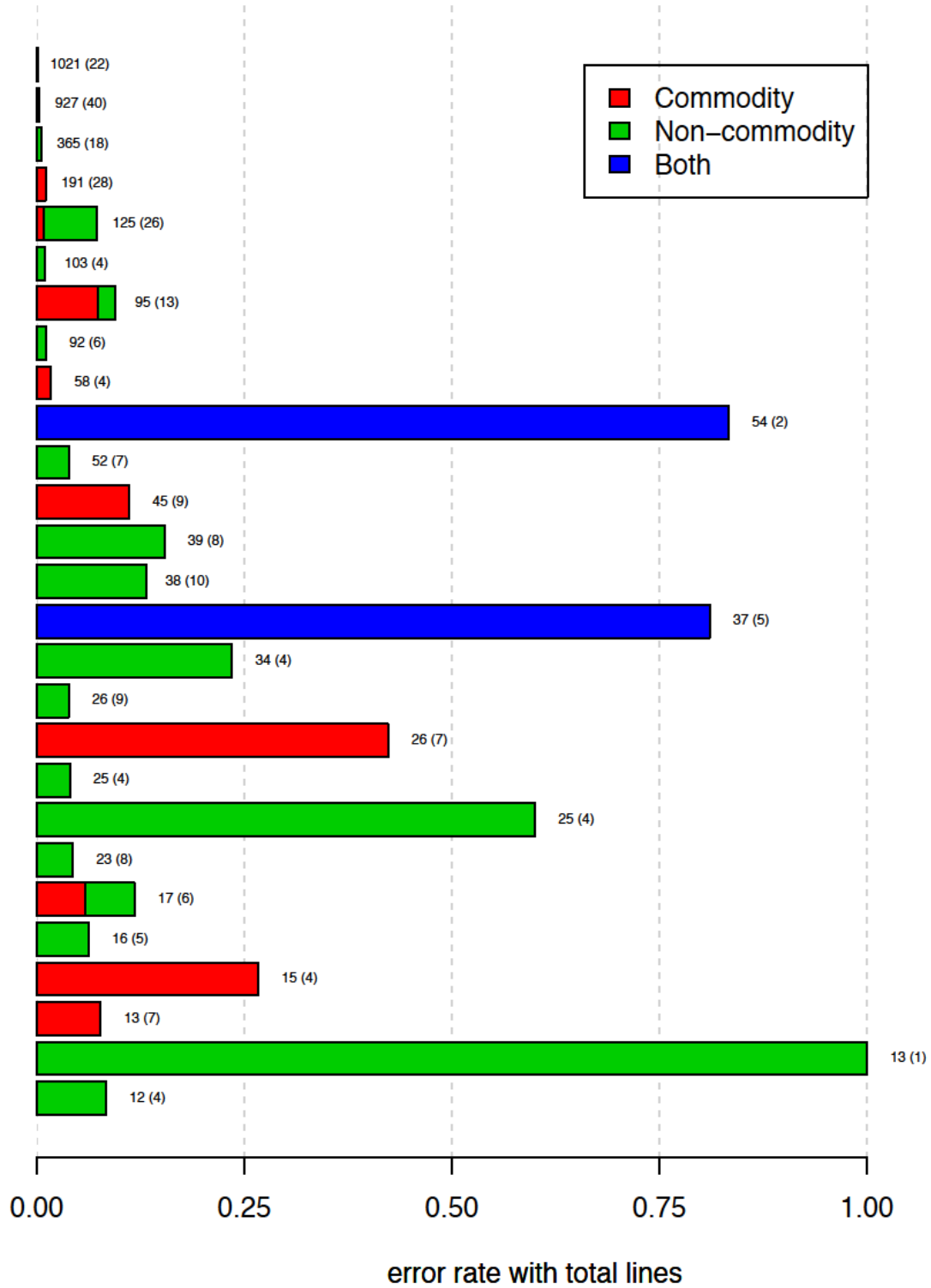
## Graphical displays of results for July 2013

The following bar charts present the failure rates by a range of factors. The labels at the right end of each bar are the numbers of lines (entries) involved.

Results by state

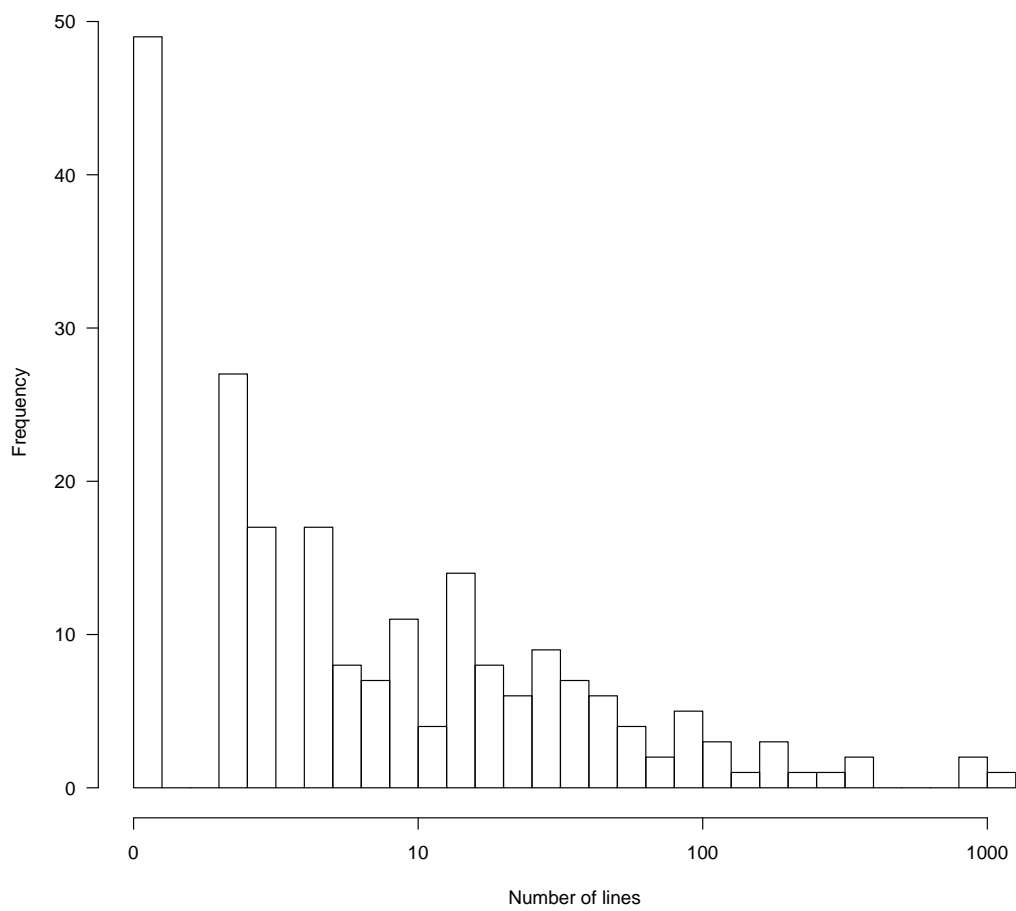


## Performance by broker

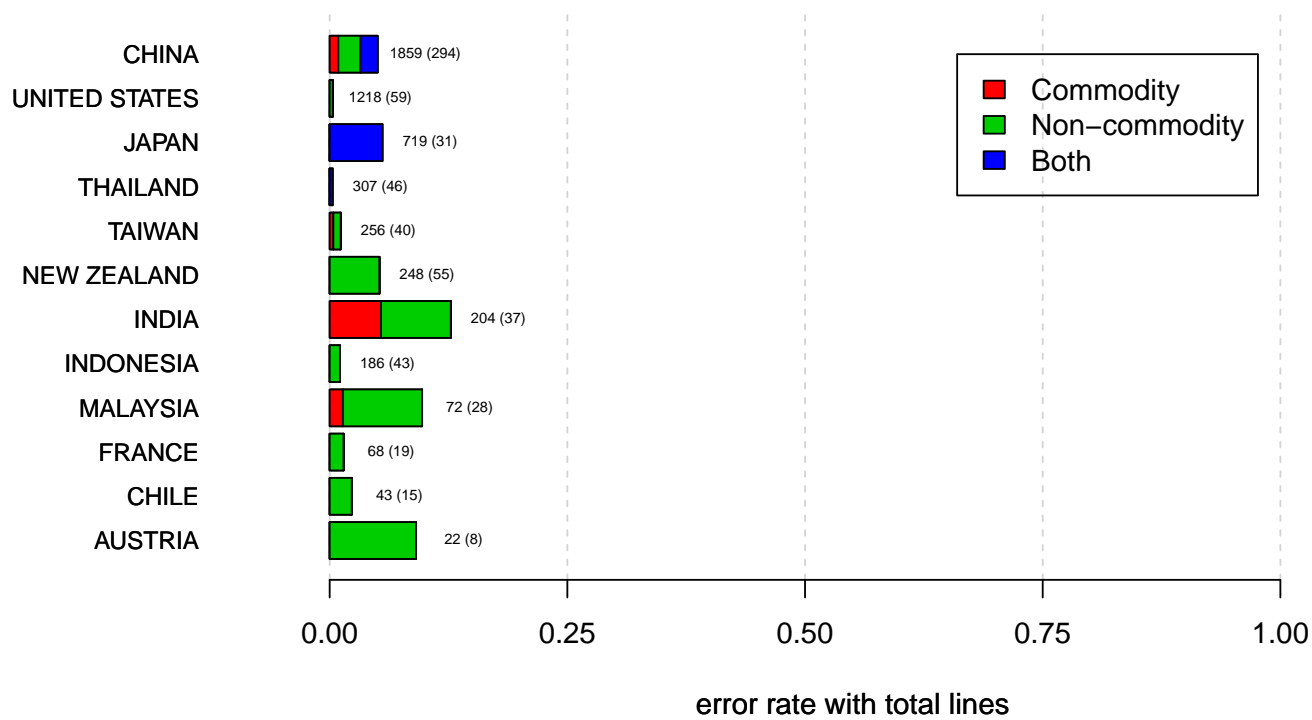


Results for the largest 27 levels with at least one failure out of a total of 38 There were a total of 177 levels with no failures

**Distribution of lines for each broker**



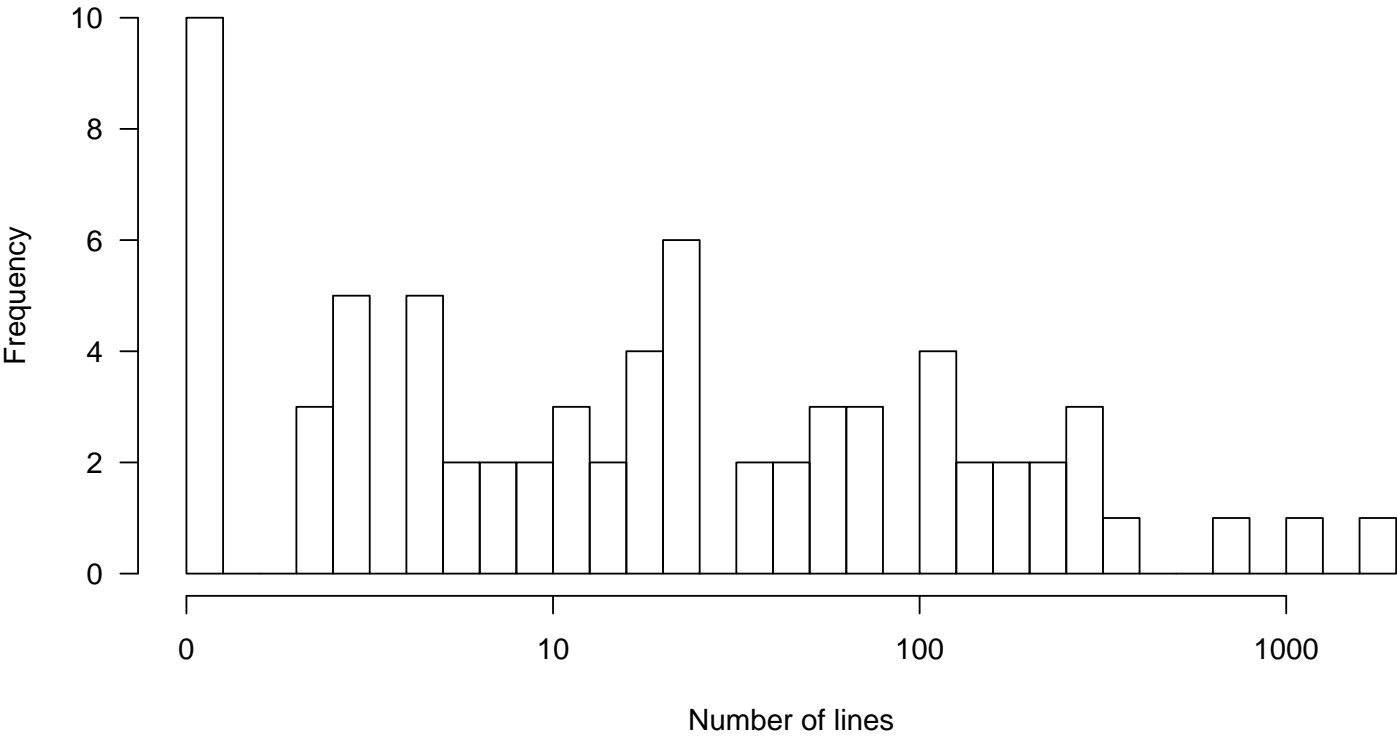
## Performance by country



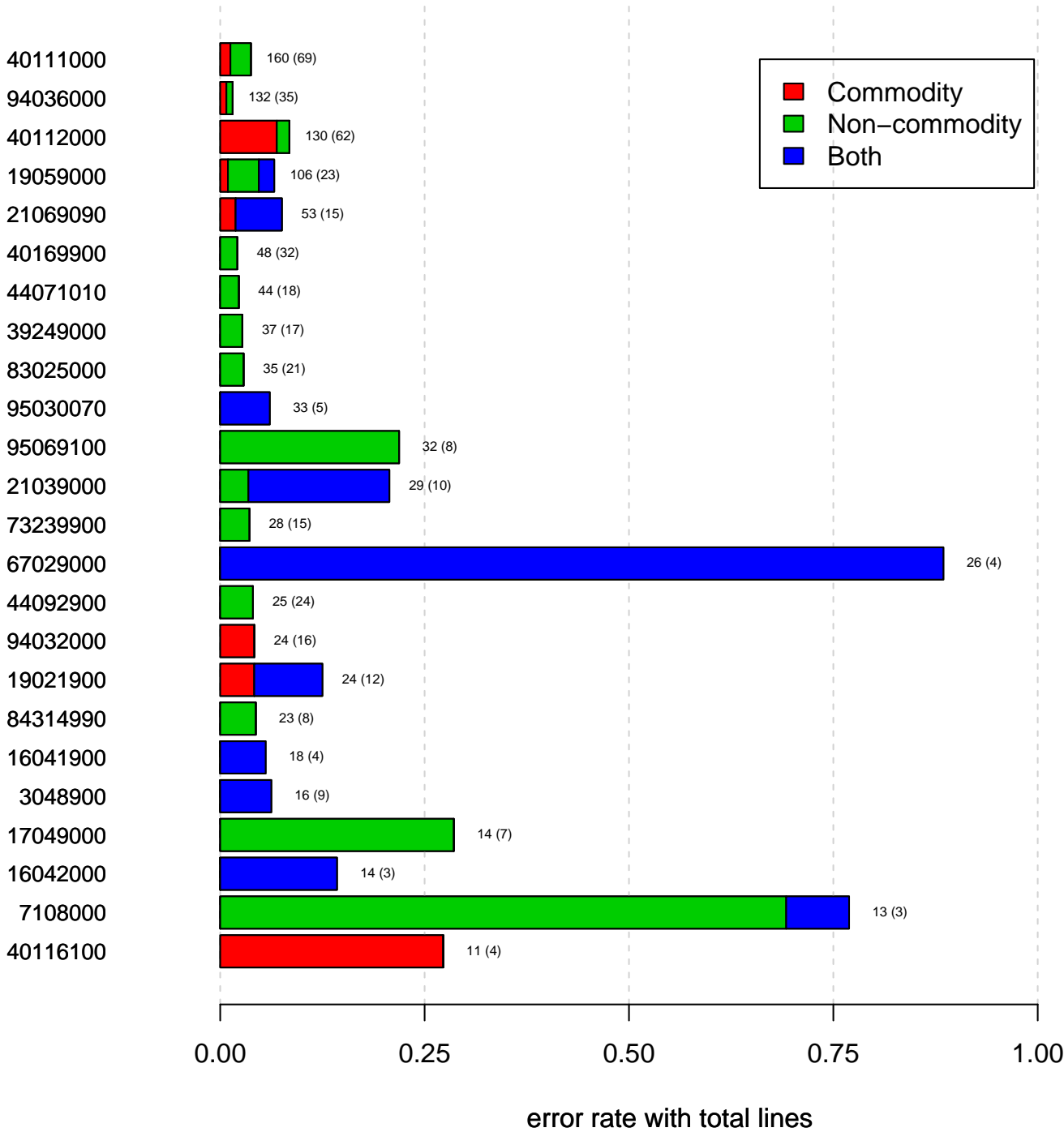
Results for the largest 12 levels with at least ten lines and one failure out of a total of 13. There were a total of 58 levels with no failures.



Distribution of lines for each country

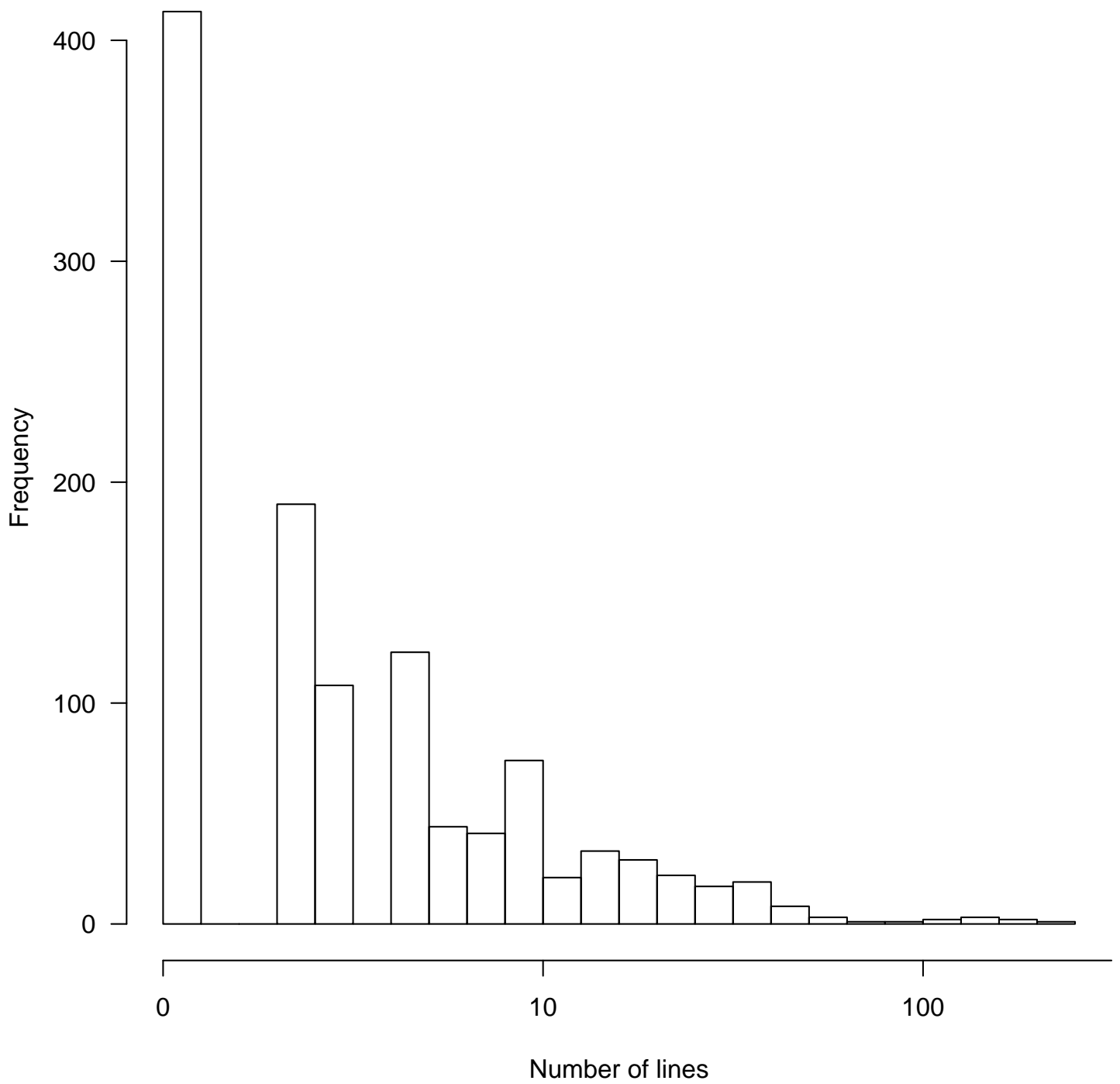


Performance by tariff

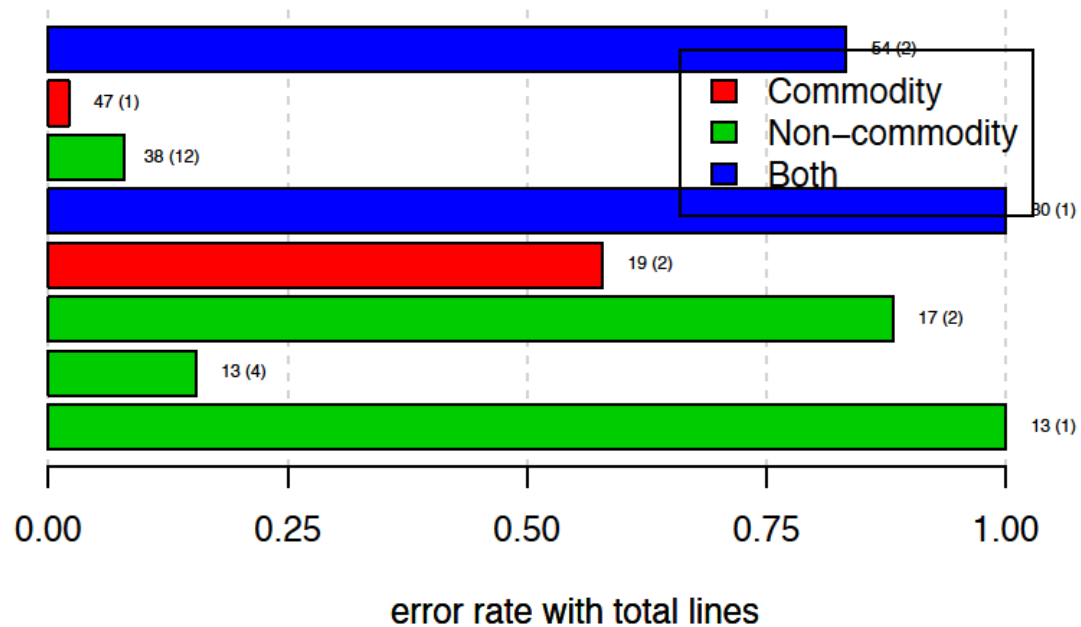


Results for the largest 24 levels with at least one failure out of a total of 100. There were a total of 1055 levels with no failures.

**Distribution of lines for each tariff**

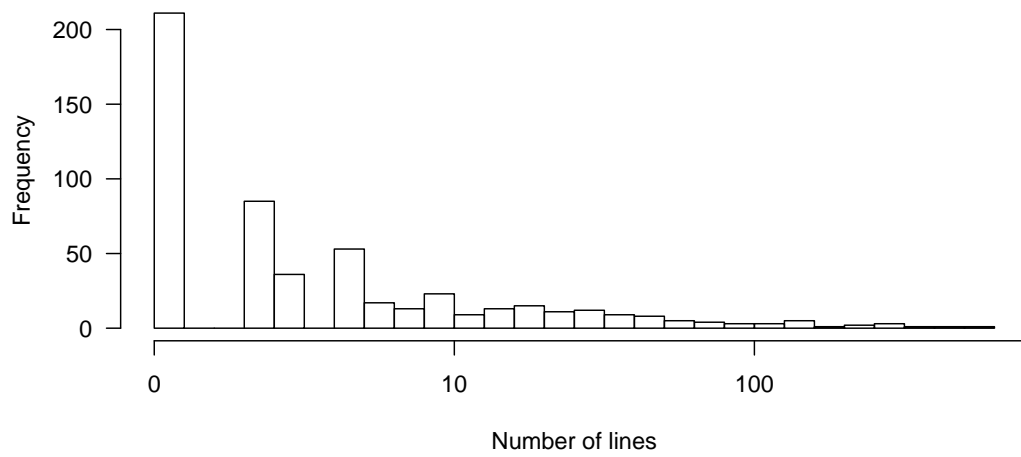


## Performance by importer code

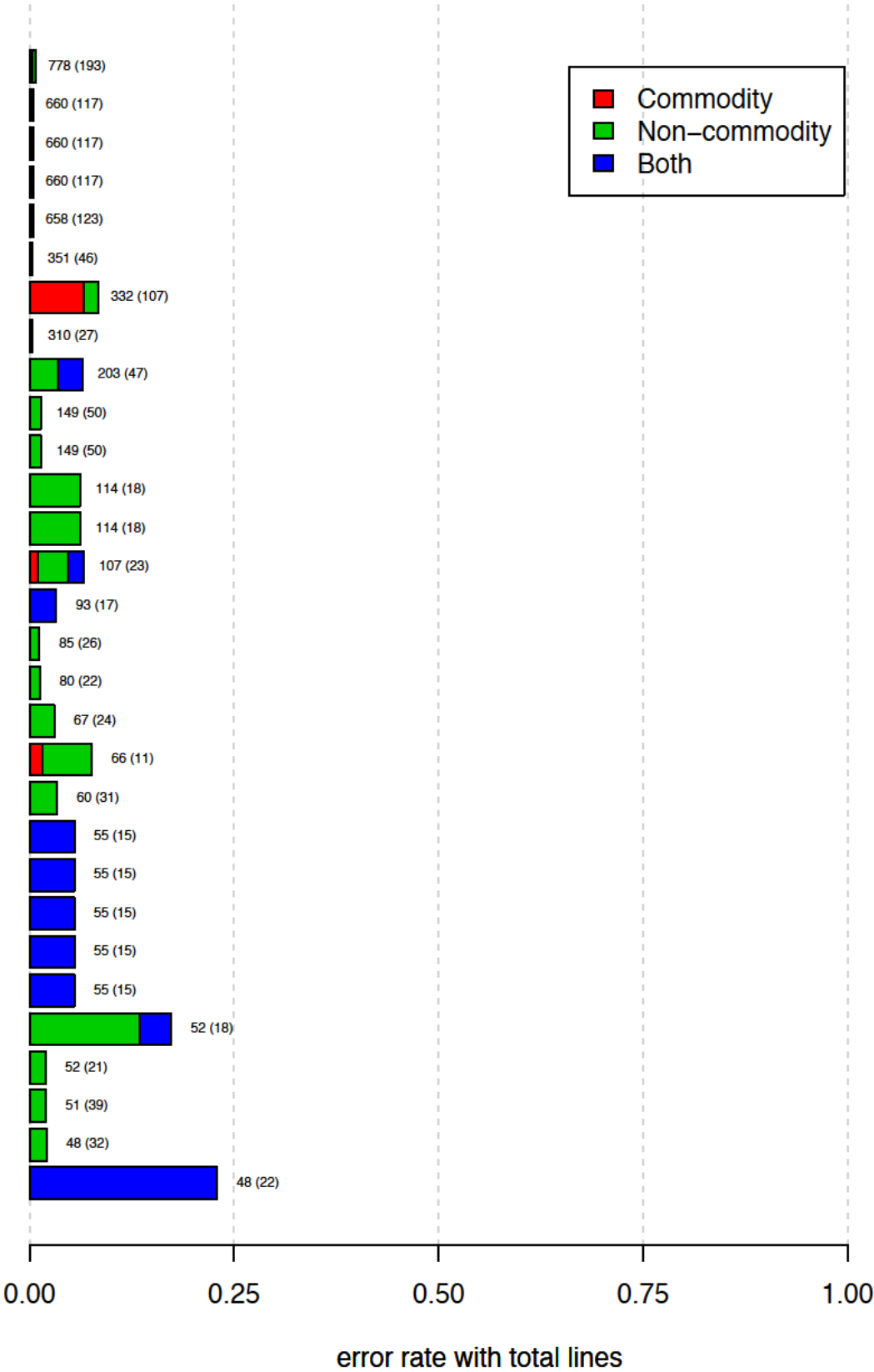


Results for the largest 8 levels with at least one failure out of a total of 42 There were a total of 502 levels with no failures

**Distribution of lines for each importer code**

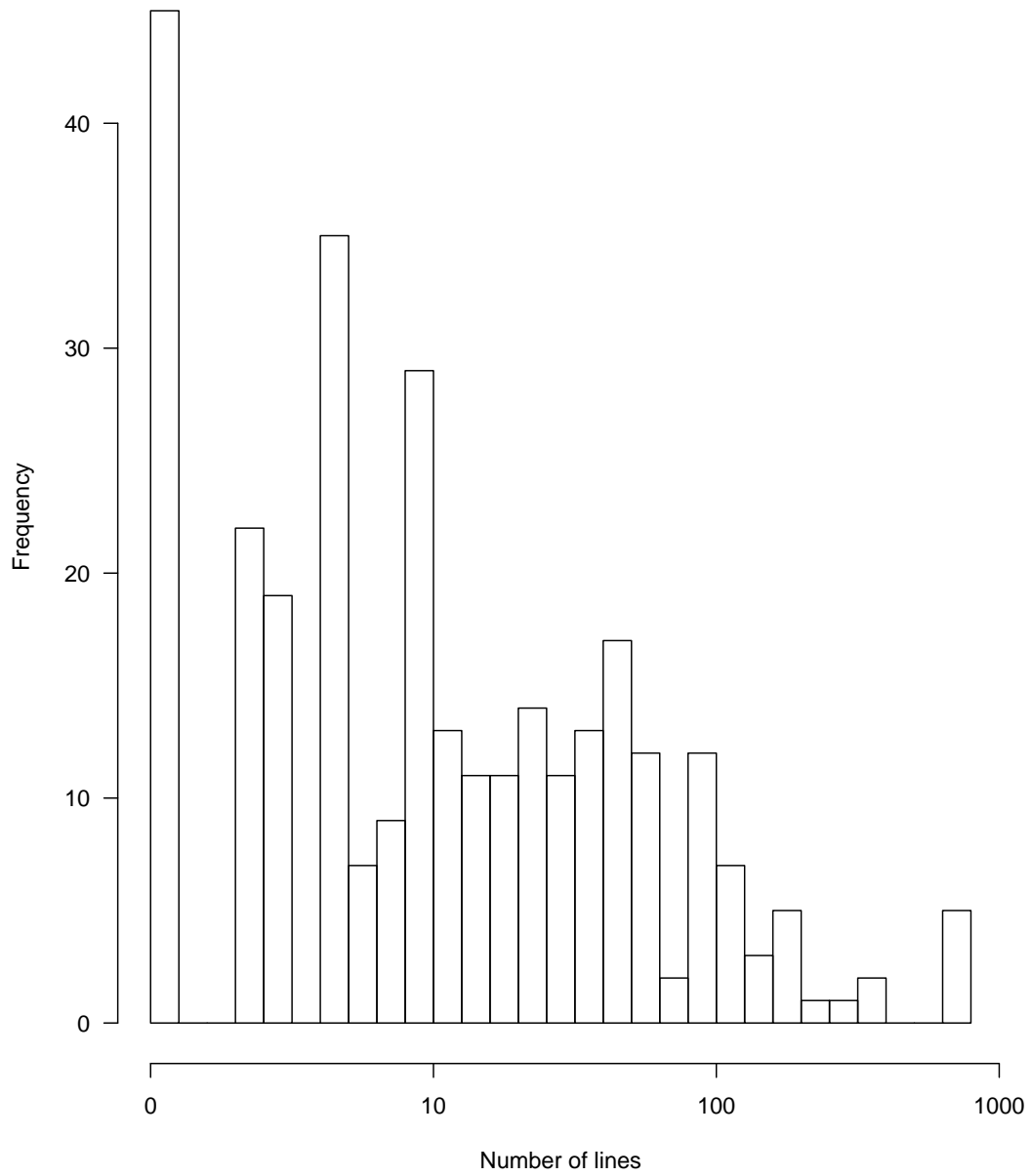


Performance by profile number

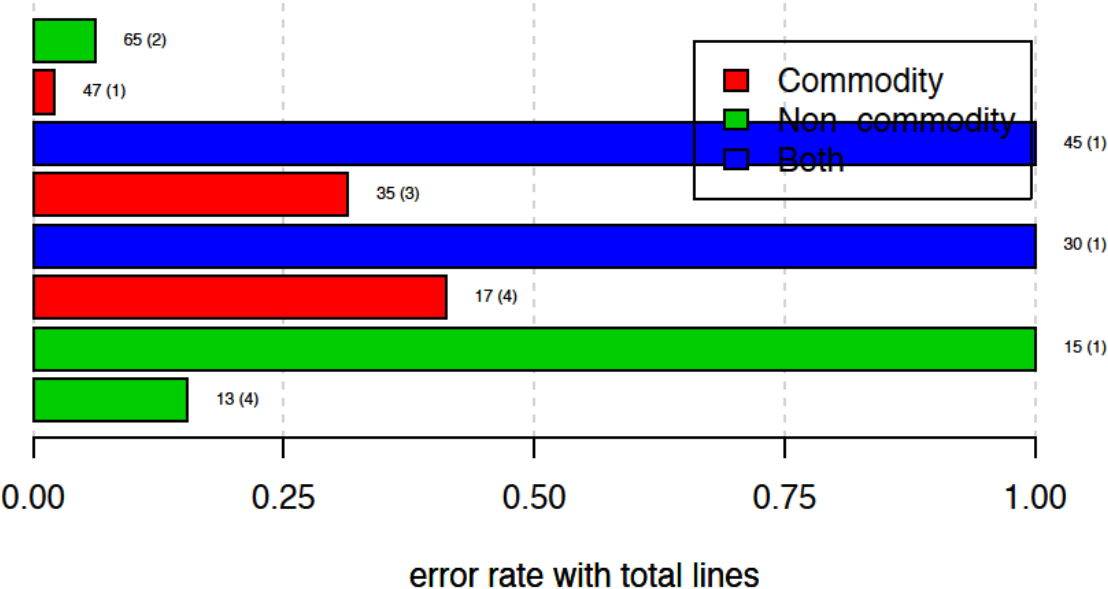


Results for the largest 30 levels with at least one failure out of a total of 97 There were a total of 209 levels with no failures

**Distribution of lines for each profile number**



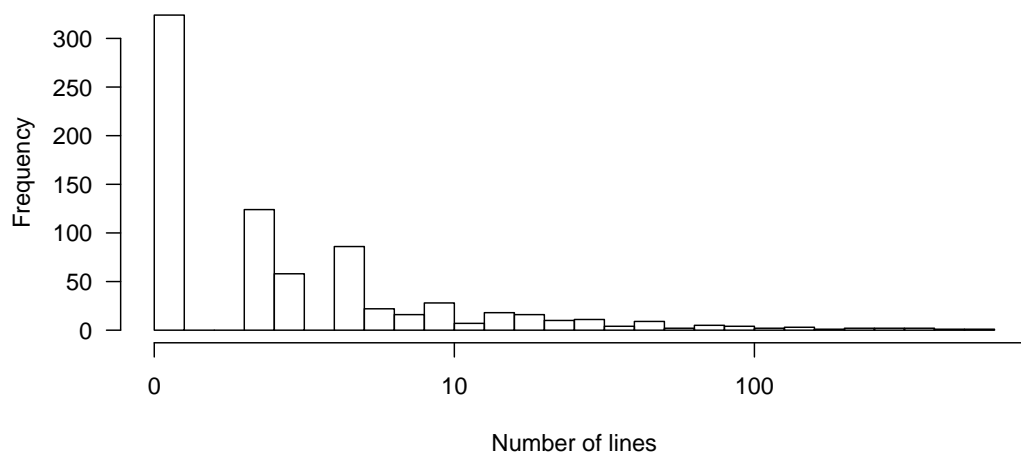
Performance by supplier code



Results for the largest 8 levels with at least one failure out of a total of 47 There were a total of 711 levels with no failures



**Distribution of lines for each supplier code**



## CCVFullScript.R

```
#####  
##Monthly CCV outputs##  
#####  
  
#Requires the following files:  
#CCVTables.r  
#CCVFlowcharts.r  
#CCVPlots.r  
  
#Requires the following libraries to be installed  
require(diagram)  
library(rJava)  
  
#SELECT MONTH#  
  
month1<-"July"  
#month1<-"August"  
  
#READ IN DATA  
#may need to change long dash to short in date column before loading  
linedata<-read.csv(paste(month1," data/CCV ",month1,  
" 2013 - line data CEBRA.csv",sep=""),stringsAsFactors=F)  
profiledata<-read.csv(paste(month1," data/CCV ",month1,  
" 2013 - line profile data CEBRA v2.csv",sep=""), stringsAsFactors=F)  
referraldata<-read.csv(paste(month1," data/CCV ",month1,  
" 2013 - entry referral data CEBRA.csv",sep=""),stringsAsFactors=F)  
  
#EXTRACT CCV ONLY (LINE DATA)  
linedataCCV<-linedata  
linedataCCV<-linedataCCV[linedataCCV$RoD=="Y",]  
linedataCCV<-linedataCCV[linedataCCV$Inspected=="Y",]  
linedataCCV<-linedataCCV[linedataCCV$Entry.Mode=="Line",]  
nline<-dim(linedataCCV)[1]  
nentry<-length(table(linedataCCV$Quarantine.Entry))  
  
#REMOVE FOOD PROFILE (PROFILE DATA)  
profiledatanofood<-profiledata[profiledata$LineProfileReason!="Food Profile",]  
  
#GENERATING PROFILE PATH DATA  
profilepath<-vector()  
for(i in 1:nline)  
{  
  entrylineID<-linedataCCV$Unique.Entry.Line.ID[i]  
  profiletemp<-profiledatanofood[  
    profiledatanofood$Unique.Entry.Line.ID==entrylineID,]  
  
  if("Y" %in% profiletemp$LineProfileReferral)  
  {  
    if("High Risk" %in% profiletemp$LineProfileReason)  
    {  
      profilepath[i]<-"high risk"  
    } else {  
      profilepath[i]<-"sometimes referred"  
    }  
  } else {  
    if("CP Profile" %in% profiletemp$LineProfileReason)  
    {
```

```

        profilepath[i]<-"downgraded profile"
    } else {
        profilepath[i]<-"not profiled"
    }
}

#GENERATING PROFILE NUMBER DATA
profileCCVmerge<-merge(profiledatanofood,
linedataCCV[,c("Unique.Entry.Line.ID","Inspection.OK.", "BRM.Where")],
by="Unique.Entry.Line.ID")

#PRODUCE TABLES
source("CCVTables.R")

#PRODUCE FLOWCHARTS
source("CCVFlowcharts.R")

#PRODUCE PLOTS
source("CCVPlots.R")

```

## CCVTables.R

```
#RUN AS PART OF CCVFullScript.R
```

```
#Produces CCV Failure Rate Tables by a range of factors
```

```
write.xlsx <- xlsx::write.xlsx
output1<-data.frame(
  Error.rate=length(which(linedataCCV$Inspection.OK=="N"))/nline,
  Com.Error.rate=length(which((linedataCCV$Inspection.OK=="N")*
    (linedataCCV$BRM.Where=="Com Fail")==1))/nline,
  NonCom.Error.rate=length(which((linedataCCV$Inspection.OK=="N")*
    (linedataCCV$BRM.Where=="Non-Com Fail")==1))/nline,
  Both.Error.rate=length(which((linedataCCV$Inspection.OK=="N")*
    (linedataCCV$BRM.Where=="Com Fail & Non-Com Fail")==1))/nline,
  Total=nline,entry.total=nentry)
rownames(output1)<-"Overall"
colnames(output1)<-c("Overall error rate", "Com error rate",
  "Non-Com error rate","Both error rate","Line total", "Entry total")
write.xlsx(output1,paste(month1," CCV results.xlsx",sep=""),
  sheetName="Overall")
```

```
#FAILURE RATE TABLE BY PROCESSING STATE
```

```
tabtemp<-cbind(prop.table(table(linedataCCV$Processing.State,
  linedataCCV$Inspection.OK.),1)[,1],
  table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
    linedataCCV$Processing.State)[1,1,]/
    table(linedataCCV$Processing.State),
  table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
    linedataCCV$Processing.State)[1,3,]/
    table(linedataCCV$Processing.State),
  table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
    linedataCCV$Processing.State)[1,2,]/
    table(linedataCCV$Processing.State),
  table(linedataCCV$Processing.State),
  rowSums((table(linedataCCV$Processing.State,
    linedataCCV$Quarantine.Entry)>0)*1))
colnames(tabtemp)<-c("Overall error rate", "Com error rate",
  "Non-Com error rate","Both error rate","Line total","Entry totals")
write.xlsx(tabtemp,paste(month1," CCV results.xlsx",sep=""),
  sheetName="Processing State", append=TRUE)
```

```
#FAILURE RATE TABLE BY COUNTRY
```

```
tabtemp<-cbind(prop.table(table(linedataCCV$Country,
  linedataCCV$Inspection.OK.),1)[,1],
  table(linedataCCV$Inspection.OK.,
    linedataCCV$BRM.Where,linedataCCV$Country)[1,1,]/
    table(linedataCCV$Country),
  table(linedataCCV$Inspection.OK.,
    linedataCCV$BRM.Where,linedataCCV$Country)[1,3,]/
    table(linedataCCV$Country),
  table(linedataCCV$Inspection.OK.,
    linedataCCV$BRM.Where,linedataCCV$Country)[1,2,]/
    table(linedataCCV$Country),
  table(linedataCCV$Country),rowSums((table(linedataCCV$Country,
    linedataCCV$Quarantine.Entry)>0)*1))
colnames(tabtemp)<-c("Overall error rate", "Com error rate",
```

```
"Non-Com error rate","Both error rate","Line total","Entry totals")
write.xlsx(tabtemp,paste(month1," CCV results.xlsx",sep=""),
sheetName="Country", append=TRUE)
```

#### #FAILURE RATE TABLE BY TARIFF

```
tabtemp<-cbind(prop.table(table(linedataCCV$Tariff,
linedataCCV$Inspection.OK.),1)[,1],
  table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,linedataCCV$Tariff)[1,1,]/
  table(linedataCCV$Tariff),
  table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,linedataCCV$Tariff)[1,3,]/
  table(linedataCCV$Tariff),
  table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,linedataCCV$Tariff)[1,2,]/
  table(linedataCCV$Tariff),
  table(linedataCCV$Tariff),rowSums((table(linedataCCV$Tariff,
linedataCCV$Quarantine.Entry)>0)*1))
colnames(tabtemp)<-c("Overall error rate", "Com error rate",
"Non-Com error rate","Both error rate","Line total","Entry totals")
ord<-order(tabtemp[,5],decreasing=T)
tabtemp<-tabtemp[ord,]
write.xlsx(tabtemp,paste(month1," CCV results.xlsx",sep=""),
sheetName="Tariff", append=TRUE)
```

#### #FAILURE RATE TABLE BY BROKER

```
tabtemp<-cbind(prop.table(table(linedataCCV$Brokeragename,
linedataCCV$Inspection.OK.),1)[,1],
  table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Brokeragename)[1,1,]/
  table(linedataCCV$Brokeragename),
  table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Brokeragename)[1,3,]/
  table(linedataCCV$Brokeragename),
  table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Brokeragename)[1,2,]/
  table(linedataCCV$Brokeragename),
  table(linedataCCV$Brokeragename),
  rowSums((table(linedataCCV$Brokeragename,
linedataCCV$Quarantine.Entry)>0)*1))
colnames(tabtemp)<-c("Overall error rate", "Com error rate",
"Non-Com error rate","Both error rate","Line total","Entry totals")
ord<-order(tabtemp[,5],decreasing=T)
tabtemp<-tabtemp[ord,]
write.xlsx(tabtemp,paste(month1," CCV results.xlsx",sep=""),
sheetName="Broker", append=TRUE)
```

#### #FAILURE RATE TABLE BY IMPORTER CODE

```
tabtemp<-cbind(prop.table(table(linedataCCV$Importer.Code,
linedataCCV$Inspection.OK.),1)[,1],
  table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Importer.Code)[1,1,]/
  table(linedataCCV$Importer.Code),
  table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Importer.Code)[1,3,]/
```

```

table(linedataCCV$Importer.Code),
table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Importer.Code)[1,2,]/
table(linedataCCV$Importer.Code),
table(linedataCCV$Importer.Code),
rowSums((table(linedataCCV$Importer.Code,
linedataCCV$Quarantine.Entry)>0)*1))
colnames(tabtemp)<-c("Overall error rate", "Com error rate",
"Non-Com error rate","Both error rate","Line total","Entry totals")
ord<-order(tabtemp[,5],decreasing=T)
tabtemp<-tabtemp[ord,]
write.xlsx(tabtemp,paste(month1," CCV results.xlsx",sep=""),
sheetName="Importer Code", append=TRUE)

```

#### #FAILURE RATE TABLE BY SUPPLIER CODE

```

tabtemp<-cbind(prop.table(table(linedataCCV$Supplier.Code,
linedataCCV$Inspection.OK.),1)[,1],
table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Supplier.Code)[1,1,]/
table(linedataCCV$Supplier.Code),
table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Supplier.Code)[1,3,]/
table(linedataCCV$Supplier.Code),
table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Supplier.Code)[1,2,]/
table(linedataCCV$Supplier.Code),
table(linedataCCV$Supplier.Code),
rowSums((table(linedataCCV$Supplier.Code,
linedataCCV$Quarantine.Entry)>0)*1))
colnames(tabtemp)<-c("Overall error rate", "Com error rate",
"Non-Com error rate","Both error rate","Line total","Entry totals")
ord<-order(tabtemp[,5],decreasing=T)
tabtemp<-tabtemp[ord,]
write.xlsx(tabtemp,paste(month1," CCV results.xlsx",sep=""),
sheetName="Supplier Code", append=TRUE)

```

#### #FAILURE RATE TABLE BY PROFILE PATH

```

tabtemp<-cbind(prop.table(table(profilepath,
linedataCCV$Inspection.OK.),1)[,1],
table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,profilepath)[1,1,]/table(profilepath),
table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,profilepath)[1,3,]/table(profilepath),
table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,profilepath)[1,2,]/table(profilepath),
table(profilepath),rowSums((table(profilepath,
linedataCCV$Quarantine.Entry)>0)*1))
colnames(tabtemp)<-c("Overall error rate", "Com error rate",
"Non-Com error rate","Both error rate","Line total","Entry totals")
write.xlsx(tabtemp,paste(month1," CCV results.xlsx",sep=""),
sheetName="Profile pathway", append=TRUE)

```

#### #FAILURE RATE TABLE BY PROFILE NUMBER

```

tabtemp<-cbind(prop.table(table(profileCCVmerge$LineProfilenum,

```

```

profileCCVmerge$Inspection.OK.),1)[,1],
  table(profileCCVmerge$Inspection.OK.,
profileCCVmerge$BRM.Where,profileCCVmerge$LineProfilenum)[1,1,]/
  table(profileCCVmerge$LineProfilenum),
  table(profileCCVmerge$Inspection.OK.,
profileCCVmerge$BRM.Where,profileCCVmerge$LineProfilenum)[1,3,]/
  table(profileCCVmerge$LineProfilenum),
  table(profileCCVmerge$Inspection.OK.,
profileCCVmerge$BRM.Where,profileCCVmerge$LineProfilenum)[1,2,]/
  table(profileCCVmerge$LineProfilenum),
  table(profileCCVmerge$LineProfilenum),
  rowSums((table(profileCCVmerge$LineProfilenum,
profileCCVmerge$Quarantine.Entry)>0)*1))
colnames(tabtemp)<-c("Overall error rate", "Com error rate",
"Non-Com error rate","Both error rate","Line total","Entry totals")
ord<-order(tabtemp[,5],decreasing=T)
tabtemp<-tabtemp[ord,]
write.xlsx(tabtemp,paste(month1," CCV results.xlsx",sep=""),
sheetName="Profile number", append=TRUE)

```

## CCVFlowcharts.R

```
#RUN AS PART OF CCVFullScript.R
```

```
#Produces CCV Flowchart for profile pathways
```

```
tabtemp<-cbind(prop.table(table(profilepath,
linedataCCV$Inspection.OK.),1)[,1],
               table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,profilepath)[1,1,]/
               table(profilepath),
               table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,profilepath)[1,2,]/
               table(profilepath),
               table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,profilepath)[1,3,]/
               table(profilepath),
               table(profilepath))
profiletable<-tabtemp
pN1<-cbind(profiletable[,2],profiletable[,2]*profiletable[,5])
pN2<-cbind(profiletable[,3],profiletable[,3]*profiletable[,5])
pN3<-cbind(profiletable[,4],profiletable[,4]*profiletable[,5])
pY<-cbind((1-profiletable[,1]),(1-profiletable[,1])*profiletable[,5])
```

```
#CREATING TRANSITION MATRIX
```

```
TM<-matrix(nrow=70, ncol=70,data=0)
TM[5,15]<-""
TM[12,21]<-""
TM[12,22]<-""
TM[13,12]<-""
TM[15,13]<- "Cleared"
TM[15,17]<- "Referred"
TM[17,18]<-""
TM[18,19]<-""
TM[17,27]<-""
TM[19,28]<-""
TM[19,29]<-""
TM[15,34]<- "Downgraded"
TM[15,36]<- "Partially"
TM[22,31]<-""
TM[22,32]<-""
TM[22,33]<-""
TM[29,38]<-""
TM[29,39]<-""
TM[29,40]<-""
TM[34,44]<-""
TM[44,53]<-""
TM[44,54]<-""
TM[36,45]<-""
TM[36,46]<-""
TM[46,47]<-""
TM[47,56]<-""
TM[47,57]<-""
TM[54,63]<-""
TM[54,64]<-""
TM[54,65]<-""
TM[57,66]<-""
TM[57,67]<-""
```



```

TM[57,68]<-"
TM=t(TM)

#PLOTING FLOWCHART (row per row)
pdf(paste(month1,"_profile_flowchart2.pdf",sep=""), width=10,height=7)
plotmat(A=TM,pos=rep(10,7),
name=c("", "", "", "", "ICS", "", "", "", "", "",
      "", "CCV\nSample", "Clear", "", "Profiling\nEngine", "",
      "Agriculture", "Release", "CCV\nSample", "",
      "Pass", "Fail", "", "", "", "", "Inspect", "Pass", "Fail", "",
      "\nCommodity", "\nBoth", "\nNon-commodity", "Clear", "",
      "Agriculture", "", "\nCommodity", "\nBoth", "\nNon-commodity",
      "", "", "", "CCV\nSample", "Inspect", "Release", "CCV\nSample",
      "", "", "",
      "", "", "Pass", "Fail", "", "Pass", "Fail", "", "", "",
      "", "", "\nCommodity", "\nBoth", "\nNon-commodity",
      "\nCommodity", "\nBoth", "\nNon-commodity", "", "")),
box.type=c("", "", "", "", "rect", "", "", "", "", "",
  "", "diamond", "rect", "", "diamond", "", "rect", "rect", "diamond", "",
  "ellipse", "ellipse", "", "", "", "", "rect", "ellipse", "ellipse", "",
  "", "", "", "rect", "", "rect", "", "", "", "",
  "", "", "", "diamond", "rect", "rect", "diamond", "", "", "",
  "", "", "ellipse", "ellipse", "", "ellipse", "ellipse", "", "", "",
  "", "", "", "", "", "", "", "", "", "", "")),
box.size=0.032*c(0,0,0,0,1.1,0,0,0,0,0,
  0,1.7,1.1,0,1.7,0,1.1,1.1,1.7,0,
  0.7,0.7,0,0,0,0,1.1,0.7,0.7,0,
  0.7,0.7,0.7,1.1,0,1.1,0,0.7,0.7,0.7,
  0,0,0,1.7,1.1,1.1,1.7,0,0,0,
  0,0,0.7,0.7,0,0.7,0.7,0,0,0,
  0,0,0.7,0.7,0.7,0.7,0.7,0.7,0,0),
box.col=c(16,16,16,16,16,16,16,16,16,16,
  16,15,16,16,15,16,16,16,15,16,
  3,2,16,16,16,16,16,3,2,16,
  2,2,2,16,16,16,16,2,2,2,
  16,16,16,15,16,16,15,16,16,16,
  16,16,3,2,16,3,2,16,16,16,
  16,16,2,2,2,2,2,2,16,16),
arr.pos = 0.6,
shadow.size=0,
curve=0,
dtext=-1,
arr.width=0,
cex=0.8,
box.cex=0.8,
main=paste(month1," CCV performance",sep="")
)
text(0.75,0.58,
labels=paste(round(100*pY[2,1],1),"% (" ,pY[2,2] ,")",sep=""),
cex=0.8)
text(0.75,0.44,
labels=paste(round(100*pN1[2,1],1),"% (" ,pN1[2,2] ,")",sep=""),
cex=0.8)
text(0.85,0.44,
labels=paste(round(100*pN2[2,1],1),"% (" ,pN2[2,2] ,")",sep=""),
cex=0.8)
text(0.95,0.44,
labels=paste(round(100*pN3[2,1],1),"% (" ,pN3[2,2] ,")",sep=""),

```

```

cex=0.8)
text(0.05,0.58,
labels=paste(round(100*pY[3,1],1),"% (" ,pY[3,2],")",sep=""),
cex=0.8)
text(0.05,0.44,
labels=paste(round(100*pN1[3,1],1),"% (" ,pN1[3,2],")",sep=""),
cex=0.8)
text(0.15,0.44,
labels=paste(round(100*pN2[3,1],1),"% (" ,pN2[3,2],")",sep=""),
cex=0.8)
text(0.25,0.44,
labels=paste(round(100*pN3[3,1],1),"% (" ,pN3[3,2],")",sep=""),
cex=0.8)
text(0.25,0.15,
labels=paste(round(100*pY[1,1],1),"% (" ,pY[1,2],")",sep=""),
cex=0.8)
text(0.25,0.01,
labels=paste(round(100*pN1[1,1],1),"% (" ,pN1[1,2],")",sep=""),
cex=0.8)
text(0.35,0.01,
labels=paste(round(100*pN2[1,1],1),"% (" ,pN2[1,2],")",sep=""),
cex=0.8)
text(0.45,0.01,
labels=paste(round(100*pN3[1,1],1),"% (" ,pN3[1,2],")",sep=""),
cex=0.8)
text(0.55,0.15,
labels=paste(round(100*pY[4,1],1),"% (" ,pY[4,2],")",sep=""),
cex=0.8)
text(0.55,0.01,
labels=paste(round(100*pN1[4,1],1),"% (" ,pN1[4,2],")",sep=""),
cex=0.8)
text(0.65,0.01,
labels=paste(round(100*pN2[4,1],1),"% (" ,pN2[4,2],")",sep=""),
cex=0.8)
text(0.75,0.01,
labels=paste(round(100*pN3[4,1],1),"% (" ,pN3[4,2],")",sep=""),
cex=0.8)
dev.off()

```

## CCVPlots.R

```
#RUN AS PART OF CCVFullScript.R
```

```
#Produces CCV Plots
```

```
#BARChart BY PROCESSING STATE
```

```
tabtemp<-cbind(prop.table(table(linedataCCV$Processing.State,
linedataCCV$Inspection.OK.),1)[,1],
               table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,linedataCCV$Processing.State)[1,1,]/
               table(linedataCCV$Processing.State,
linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,linedataCCV$Processing.State)[1,3,]/
               table(linedataCCV$Processing.State,
linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,linedataCCV$Processing.State)[1,2,]/t
               able(linedataCCV$Processing.State,
table(linedataCCV$Processing.State,
rowSums((table(linedataCCV$Processing.State,
linedataCCV$Quarantine.Entry)>0)*1))

data1<-tabtemp
ord<-order(data1[,5])
data1<-data1[ord,]
n<-dim(data1)[1]
pdf(paste("../graphics/",month1,"_State_barplot2.pdf",sep=""),
width=8,height=.25*n+2)
par(mar=c(5,7,4,0))
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),horiz=T,
xlab="CCV non-compliance rate with total lines (entries)",las=1,
cex.names=0.8, main="Results by state",xlim=c(-0.1,1.2),
axes=F, col=c(7,4,3))
abline(v=c(0,0.25,0.5,0.75,1),lty=2,col="light grey")
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="CCV non-compliance rate with total lines (entries)",
las=1,cex.names=0.8,main="Results by state",xlim=c(-0.1,1.2),
axes=F, add=T, col=c(7,4,3))
legend(y=max(Graph),x=0.66,c("Commodity","Non-commodity","Both"),
fill=c(7,4,3), cex=0.8)
text(data1[,1]+0.05,Graph,
paste(data1[,5]," (" ,data1[,6],")",sep=""),
cex=0.5)
axis(side = 1, at = c(0,0.25,0.5,0.75,1))
dev.off()
```

```
#BARChart BY COUNTRY
```

```
tabtemp<-cbind(prop.table(table(linedataCCV$Country,
linedataCCV$Inspection.OK.),1)[,1],
               table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,linedataCCV$Country)[1,1,]/
               table(linedataCCV$Country,
linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,linedataCCV$Country)[1,3,]/
               table(linedataCCV$Country,
table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,linedataCCV$Country)[1,2,]/
```

```

        table(linedataCCV$Country),
        table(linedataCCV$Country),rowSums((table(linedataCCV$Country,
        linedataCCV$Quarantine.Entry)>0)*1))
n1<-dim(tabtemp)[1] #number of levels
tabtemp1<-tabtemp
tabtemp<-tabtemp[tabtemp[,1]>0,]
n2<-dim(tabtemp)[1] #number of non-zero levels
n4<-n1-n2 #number zero excluded
cutoff<-max(sort(tabtemp[,5],decreasing=T)[30],10,na.rm=T)
data1<-tabtemp[tabtemp[,5]>=cutoff,]
ord<-order(data1[,5])
data1<-data1[ord,]
n<-dim(data1)[1]
n3<-n2-n #number non-zero excluded
pdf(paste("../graphics/",month1,"_Country_barplot2.pdf",sep=""),
width=9,height=.25*n+2)
par(mar=c(5,10,4,0))
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="error rate with total lines",las=1,cex.names=0.8,
main="Performance by country",xlim=c(-0.1,1.2),axes=F, col=2:4)
abline(v=c(0,0.25,0.5,0.75,1),lty=2,col="light grey")
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="error rate with total lines",las=1,cex.names=0.8,
main="Performance by country",xlim=c(-0.1,1.2),axes=F, add=T,
col=2:4)
legend(y=max(Graph),x=0.66,c("Commodity","Non-commodity","Both"),
fill=2:4)
text(data1[,1]+0.05,Graph,paste(data1[,5]," (" ,data1[,6],")",sep=""),
cex=0.5)
axis(side = 1, at = c(0,0.25,0.5,0.75,1))
mtext(paste("Results for the largest ",n,
" levels with at least ten lines and one failure out of a total of ",
n2,". There were a total of ",
n4, " levels with no failures.",sep=""),side=1,cex=0.6, line=4)
par(mar=c(5,5,3,3))
hist(log10(tabtemp1[,5]),breaks=30,
xlab="Number of lines", main="Distribution of lines for each country",
las=1, xaxt='n')
axis(side =1, at=c(0,1,2,3),labels=c(0,10,100,1000))
dev.off()

```

#### #BARCHART BY TARIFF

```

tabtemp<-cbind(prop.table(table(linedataCCV$Tariff,
linedataCCV$Inspection.OK.),1)[,1],
        table(linedataCCV$Inspection.OK.,
        linedataCCV$BRM.Where,linedataCCV$Tariff)[1,1,]/
        table(linedataCCV$Tariff),
        table(linedataCCV$Inspection.OK.,
        linedataCCV$BRM.Where,linedataCCV$Tariff)[1,3,]/
        table(linedataCCV$Tariff),
        table(linedataCCV$Inspection.OK.,
        linedataCCV$BRM.Where,linedataCCV$Tariff)[1,2,]/
        table(linedataCCV$Tariff),
        table(linedataCCV$Tariff),
        rowSums((table(linedataCCV$Tariff,
        linedataCCV$Quarantine.Entry)>0)*1))
tabtemp1<-tabtemp

```

```

n1<-dim(tabtemp)[1] #number of levels
tabtemp<-tabtemp[tabtemp[,1]>0,]
n2<-dim(tabtemp)[1] #number of non-zero levels
n4<-n1-n2 #number zero excluded
cutoff<-max(sort(tabtemp[,5],decreasing=T)[30],10,na.rm=T)
data1<-tabtemp[tabtemp[,5]>=cutoff,]
ord<-order(data1[,5])
data1<-data1[ord,]
n<-dim(data1)[1]
n3<-n2-n #number non-zero excluded
pdf(paste("../graphics/",month1,"_Tariff_barplot2.pdf",sep=""),
width=8,height=.25*n+2)
par(mar=c(5,7,4,0))
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="error rate with total lines",las=1,cex.names=0.8,
main="Performance by tariff",xlim=c(-0.1,1.2),axes=F, col=2:4)
abline(v=c(0,0.25,0.5,0.75,1),lty=2,col="light grey")
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="error rate with total lines",las=1,cex.names=0.8,
main="Performance by tariff",xlim=c(-0.1,1.2),axes=F, add=T, col=2:4)
legend(y=max(Graph),x=0.66,c("Commodity","Non-commodity","Both"),
fill=2:4)
text(data1[,1]+0.05,Graph,paste(data1[,5]," (",data1[,6],")",sep=""),
cex=0.5)
axis(side = 1, at = c(0,0.25,0.5,0.75,1))
mtext(paste("Results for the largest ",n,
" levels with at least one failure out of a total of ",
n2,". There were a total of ",
n4, " levels with no failures.",sep=""),side=1,cex=0.6,
line=4)
par(mar=c(5,5,3,3))
hist(log10(tabtemp[,5]),breaks=30,
xlab="Number of lines", main="Distribution of lines for each tariff",
las=1, xaxt='n')
axis(side =1, at=c(0,1,2,3),labels=c(0,10,100,1000))
dev.off()

#BARCHART BY BROKER

tabtemp<-cbind(prop.table(table(linedataCCV$Brokeragename,
linedataCCV$Inspection.OK.),1)[,1],
table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Brokeragename)[1,1,]/
table(linedataCCV$Brokeragename),
table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Brokeragename)[1,3,]/
table(linedataCCV$Brokeragename),
table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Brokeragename)[1,2,]/
table(linedataCCV$Brokeragename),
table(linedataCCV$Brokeragename),
rowSums((table(linedataCCV$Brokeragename,
linedataCCV$Quarantine.Entry)>0)*1))
tabtemp1<-tabtemp
n1<-dim(tabtemp)[1] #number of levels
tabtemp<-tabtemp[tabtemp[,1]>0,]
n2<-dim(tabtemp)[1] #number of non-zero levels
n4<-n1-n2 #number zero excluded

```

```

cutoff<-max(sort(tabtemp[,5],decreasing=T)[30],10,na.rm=T)
data1<-tabtemp[tabtemp[,5]>=cutoff,]
ord<-order(data1[,5])
data1<-data1[ord,]
n<-dim(data1)[1]
n3<-n2-n #number non-zero excluded
pdf(paste("../graphics/",month1,"_Broker_barplot2.pdf",sep=""),
width=10,height=.25*n+2)
par(mar=c(5,20,4,0))
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="error rate with total lines",las=1,cex.names=0.75,
main="Performance by broker",xlim=c(0,1.2),axes=F, col=2:4)
abline(v=c(0,0.25,0.5,0.75,1),lty=2,col="light grey")
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="error rate with total lines",las=1,cex.names=0.75,
main="Performance by broker",xlim=c(0,1.2),axes=F, add=T, col=2:4)
legend(y=max(Graph),x=0.66,c("Commodity","Non-commodity","Both"),
fill=2:4)
text(data1[,1]+0.05,Graph,paste(data1[,5]," (",data1[,6],")",sep=""),
cex=0.5)
axis(side = 1, at = c(0,0.25,0.5,0.75,1))
mtext(paste("Results for the largest ",n,
" levels with at least one failure out of a total of ",n2,
". There were a total of ", n4,
" levels with no failures.",sep=""),side=1,cex=0.6, line=4)
par(mar=c(5,5,3,3))
hist(log10(tabtemp[,5]),breaks=30,
xlab="Number of lines", main="Distribution of lines for each broker",
las=1, xaxt='n')
axis(side =1, at=c(0,1,2,3),labels=c(0,10,100,1000))
dev.off()

```

#### #BARCHART BY IMPORTER CODE

```

tabtemp<-cbind(prop.table(table(linedataCCV$Importer.Code,
linedataCCV$Inspection.OK.),1)[,1],
table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,linedataCCV$Importer.Code)[1,1,]/
table(linedataCCV$Importer.Code),
table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,linedataCCV$Importer.Code)[1,3,]/
table(linedataCCV$Importer.Code),
table(linedataCCV$Inspection.OK.,
linedataCCV$BRM.Where,linedataCCV$Importer.Code)[1,2,]/
table(linedataCCV$Importer.Code),
table(linedataCCV$Importer.Code),
rowSums((table(linedataCCV$Importer.Code,
linedataCCV$Quarantine.Entry)>0)*1))
tabtemp1<-tabtemp
n1<-dim(tabtemp)[1] #number of levels
tabtemp<-tabtemp[tabtemp[,1]>0,]
n2<-dim(tabtemp)[1] #number of non-zero levels
n4<-n1-n2 #number zero excluded
cutoff<-max(sort(tabtemp[,5],decreasing=T)[30],10,na.rm=T)
data1<-tabtemp[tabtemp[,5]>=cutoff,]
ord<-order(data1[,5])
data1<-data1[ord,]
n<-dim(data1)[1]

```

```

n3<-n2-n #number non-zero excluded
pdf(paste("../graphics/",month1,"_Importer_barplot2.pdf",sep=""),
width=8,height=.25*n+2)
par(mar=c(5,10,4,0))
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="error rate with total lines",las=1,cex.names=0.8,
main="Performance by importer code",xlim=c(-0.1,1.2),axes=F, col=2:4)
abline(v=c(0,0.25,0.5,0.75,1),lty=2,col="light grey")
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="error rate with total lines",las=1,cex.names=0.8,
main="Performance by importer code",xlim=c(-0.1,1.2),axes=F, add=T,
col=2:4)
legend(y=max(Graph),x=0.66,c("Commodity","Non-commodity","Both"),
fill=2:4)
text(data1[,1]+0.05,Graph,paste(data1[,5],"(",data1[,6],")",sep=""),
cex=0.5)
axis(side = 1, at = c(0,0.25,0.5,0.75,1))
mtext(paste("Results for the largest ",n,
" levels with at least one failure out of a total of ",n2,
". There were a total of ", n4, " levels with no failures.",sep=""),
side=1,cex=0.6, line=4)
par(mar=c(5,5,3,3))
hist(log10(tabtemp1[,5]),breaks=30, xlab="Number of lines",
main="Distribution of lines for each importer code",
las=1, xaxt='n')
axis(side =1, at=c(0,1,2,3),labels=c(0,10,100,1000))
dev.off()

#BARCHART BY SUPPLIER CODE
tabtemp<-cbind(prop.table(table(linedataCCV$Supplier.Code,
linedataCCV$Inspection.OK.),1)[,1],
table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Supplier.Code)[1,1,]/table(linedataCCV$Supplier.Code),
table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Supplier.Code)[1,3,]/table(linedataCCV$Supplier.Code),
table(linedataCCV$Inspection.OK.,linedataCCV$BRM.Where,
linedataCCV$Supplier.Code)[1,2,]/table(linedataCCV$Supplier.Code),
table(linedataCCV$Supplier.Code),
rowSums((table(linedataCCV$Supplier.Code,
linedataCCV$Quarantine.Entry)>0)*1))
tabtemp1<-tabtemp
n1<-dim(tabtemp)[1] #number of levels
tabtemp<-tabtemp[tabtemp[,1]>0,]
n2<-dim(tabtemp)[1] #number of non-zero levels
n4<-n1-n2 #number zero excluded
cutoff<-max(sort(tabtemp[,5],decreasing=T)[30],10,na.rm=T)
data1<-tabtemp[tabtemp[,5]>=cutoff,]
ord<-order(data1[,5])
data1<-data1[ord,]
n<-dim(data1)[1]
n3<-n2-n #number non-zero excluded
pdf(paste("../graphics/",month1,"_Supplier_barplot2.pdf",sep=""),
width=8,height=.25*n+2)
par(mar=c(5,7,4,0))
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="error rate with total lines",las=1,cex.names=0.8,
main="Performance by supplier code",xlim=c(-0.1,1.2),axes=F, col=2:4)
abline(v=c(0,0.25,0.5,0.75,1),lty=2,col="light grey")

```

```

Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="error rate with total lines",las=1,cex.names=0.8,
main="Performance by supplier code",xlim=c(-0.1,1.2),axes=F, add=T,
col=2:4)
legend(y=max(Graph),x=0.66,c("Commodity","Non-commodity","Both"),
fill=2:4)
text(data1[,1]+0.05,Graph,paste(data1[,5]," (" ,data1[,6],")",sep=""),
cex=0.5)
axis(side = 1, at = c(0,0.25,0.5,0.75,1))
mtext(paste("Results for the largest ",n,
" levels with at least one failure out of a total of ",n2,
". There were a total of ", n4,
" levels with no failures.",sep=""),side=1,cex=0.6, line=4)
par(mar=c(5,5,3,3))
hist(log10(tabtemp1[,5]),breaks=30, xlab="Number of lines",
main="Distribution of lines for each supplier code",
las=1, xaxt='n')
axis(side =1, at=c(0,1,2,3),labels=c(0,10,100,1000))
dev.off()

#BARCHART BY PROFILE NUMBER
tabtemp<-cbind(prop.table(table(profileCCVmerge$LineProfilenum,
profileCCVmerge$Inspection.OK.),1)[,1],
table(profileCCVmerge$Inspection.OK.,
profileCCVmerge$BRM.Where,profileCCVmerge$LineProfilenum)[1,1,]/
table(profileCCVmerge$LineProfilenum),
table(profileCCVmerge$Inspection.OK.,
profileCCVmerge$BRM.Where,profileCCVmerge$LineProfilenum)[1,3,]/
table(profileCCVmerge$LineProfilenum),
table(profileCCVmerge$Inspection.OK.,
profileCCVmerge$BRM.Where,profileCCVmerge$LineProfilenum)[1,2,]/
table(profileCCVmerge$LineProfilenum),
table(profileCCVmerge$LineProfilenum),
rowSums((table(profileCCVmerge$LineProfilenum,
profileCCVmerge$Quarantine.Entry)>0)*1))
tabtemp1<-tabtemp
n1<-dim(tabtemp)[1] #number of levels
tabtemp<-tabtemp[tabtemp[,1]>0,]
n2<-dim(tabtemp)[1] #number of non-zero levels
n4<-n1-n2 #number zero excluded
cutoff<-max(sort(tabtemp[,5],decreasing=T)[30],10,na.rm=T)
data1<-tabtemp[tabtemp[,5]>=cutoff,]
ord<-order(data1[,5])
data1<-data1[ord,]
n<-dim(data1)[1]
n3<-n2-n #number non-zero excluded
pdf(paste("../graphics/",month1,"_Profiles_barplot2.pdf",sep=""),
width=8,height=.25*n+2)
par(mar=c(5,7,4,0))
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="error rate with total lines",las=1,cex.names=0.8,
main="Performance by profile number",xlim=c(-0.1,1.2),axes=F, col=2:4)
abline(v=c(0,0.25,0.5,0.75,1),lty=2,col="light grey")
Graph<-barplot(t(data1[,2:4]),names.arg=rownames(data1),
horiz=T,xlab="error rate with total lines",las=1,cex.names=0.8,
main="Performance by profile number",xlim=c(-0.1,1.2),axes=F, add=T,
col=2:4)

```



```

legend(y=max(Graph),x=0.66,c("Commodity","Non-commodity","Both"),
fill=2:4)
text(data1[,1]+0.05,Graph,paste(data1[,5],"(",data1[,6],")",sep=""),
cex=0.5)
axis(side = 1, at = c(0,0.25,0.5,0.75,1))
mtext(paste("Results for the largest ",n,
" levels with at least one failure out of a total of ",n2,
". There were a total of ", n4,
" levels with no failures.",sep=""),side=1,cex=0.6, line=4)
par(mar=c(5,5,3,3))
hist(log10(tabtemp1[,5]),breaks=30, xlab="Number of lines",
main="Distribution of lines for each profile number",
las=1, xaxt='n')
axis(side =1, at=c(0,1,2,3),labels=c(0,10,100,1000))
dev.off()

```

## GlobalSummaries.R

```
#####
#### global summaries ####
#####

#SELECT MONTH#

month1<-"July"
#month1<-"August"

#READ IN DATA
linedata<-read.csv(paste(month1," data/CCV ",
month1," 2013 - line data CEBRA.csv",sep=""),
stringsAsFactors=F) #change long dash to short in column 5 before loading
profiledata<-read.csv(paste(month1," data/CCV ",
month1," 2013 - line profile data CEBRA v2.csv",sep=""),
stringsAsFactors=F)
referraldata<-read.csv(paste(month1," data/CCV ",
month1," 2013 - entry referral data CEBRA.csv",sep=""),
stringsAsFactors=F)
nline<-dim(linedata)[1]
nentry<-length(table(linedata$Quarantine.Entry))

#SELECT COLUMNS
data0<-profiledata[,
c("Unique.Entry.Line.ID","Quarantine.Entry","LineProfileReferral")]
data1<-referraldata[,c("Quarantine.Entry","EntryProfileReferral")]

#EXTRACT REFERRAL INFORMATION FOR LINES AND ENTRIES
linereferral<-entryreferral<-vector()
for(i in 1:nline)
{
  ID1<-linedata$Unique.Entry.Line.ID[i]
  ID2<-linedata$Quarantine.Entry[i]
  temp<-data0[data0$Unique.Entry.Line.ID==ID1,]
  temp2<-data1[data1$Quarantine.Entry==ID2,]

  if("Y" %in% temp$LineProfileReferral)
  {
    linereferral[i]<-"Y"
  } else {
    linereferral[i]<-"N"
  }
  if("Y" %in% temp2$EntryProfileReferral)
  {
    entryreferral[i]<-"Y"
  } else {
    entryreferral[i]<-"N"
  }
}

#MERGE PROFILE AND LINE DATA
data2<-data.frame(linedata[,c("Unique.Entry.Line.ID",
"Quarantine.Entry", "RoD","Inspected","Inspection.OK.")],
linereferral,entryreferral)

write.csv(data2,paste(month1,"_fulldata.csv",sep=""))
```

```

#NO REFERRALS
data2.b<-data2[(data2$linereferral=="N") &
(data2$entryreferral=="N") & (data2$RoD=="Y") &
(data2$Inspected=="Y") & (data2$Inspection.OK=="N") ,]
data2.c<-data2[(data2$linereferral=="N") &
(data2$entryreferral=="N") & (data2$RoD=="Y") &
(data2$Inspected=="Y") & (data2$Inspection.OK=="Y") ,]

#ENTRY REFERRALS
data2.d<-data2[(data2$linereferral=="N") &
(data2$entryreferral=="Y") & (data2$RoD=="N") &
(data2$Inspected=="Y") & (data2$Inspection.OK=="N") ,]
data2.e<-data2[(data2$linereferral=="N") &
(data2$entryreferral=="Y") & (data2$RoD=="N") &
(data2$Inspected=="Y") & (data2$Inspection.OK=="Y") ,]
data2.f<-data2[(data2$linereferral=="N") &
(data2$entryreferral=="Y") & (data2$RoD=="Y") &
(data2$Inspected=="N") ,]
data2.g<-data2[(data2$linereferral=="N") &
(data2$entryreferral=="Y") & (data2$RoD=="Y") &
(data2$Inspected=="Y") & (data2$Inspection.OK=="N") ,]
data2.h<-data2[(data2$linereferral=="N") &
(data2$entryreferral=="Y") & (data2$RoD=="Y") &
(data2$Inspected=="Y") & (data2$Inspection.OK=="Y") ,]

#LINE REFERRALS
data2.i<-data2[(data2$linereferral=="Y") &
(data2$RoD=="N") & (data2$Inspected=="Y") &
(data2$Inspection.OK=="N") ,]
data2.j<-data2[(data2$linereferral=="Y") &
(data2$RoD=="N") & (data2$Inspected=="Y") &
(data2$Inspection.OK=="Y") ,]
data2.k<-data2[(data2$linereferral=="Y") &
(data2$RoD=="Y") & (data2$Inspected=="N") ,]
data2.l<-data2[(data2$linereferral=="Y") &
(data2$RoD=="Y") & (data2$Inspected=="Y") &
(data2$Inspection.OK=="N") ,]
data2.m<-data2[(data2$linereferral=="Y") &
(data2$RoD=="Y") & (data2$Inspected=="Y") &
(data2$Inspection.OK=="Y") ,]

#COUNTING VALUE AT EACH NODE
nodes<-c("a","b","c","d","e","f","g","h","i","j","k","l","m")
line.n<-vector()
entry.n<-vector()
for(i in 2:13)
{
line.n[i]<-dim(get(paste("data2.",nodes[i],sep="")))[1]
entry.n[i]<-length(table(get(paste("data2.",nodes[i],sep=""))[,2]))
}

#A POSSIBLE USEFULNESS MEASURE
OR<-line.n[nodes=="l"]*line.n[nodes=="c"] /
(line.n[nodes=="m"]*line.n[nodes=="b"])

selogOR<-sqrt(1/line.n[nodes=="l"]+1/line.n[nodes=="c"]+
1/line.n[nodes=="m"]+1/line.n[nodes=="b"])

```

```
CIOR<-c(exp(log(OR)-1.96*selogOR),exp(log(OR)+1.96*selogOR))
```