cebra
Centre of Excellence for Biosecurity Risk Analysis

Report Cover Page

CEBRA Project		
170618		
Title		
Optimising New Zealand's marine biosecurity surveillance programme: Comparing the vessel traffic of different data sources using the Levenshtein distance.		
Author(s) / Address (es)		
Rezvan Hatami (CEBRA), Tracey Hollings (CEBRA), Andrew Robinson (CEBRA), Graeme Inglis (NIWA), Abraham Growcott (MPI), Daniel Kluza (MPI), Catherine Lubarsky (MPI)		
Material Type and Status (Internal draft, Final Technical or Project report, Manuscript, Manual, Software)		
Interim report for February milestone – MPI/Lloyd's data comparison work		
Summary		
CEBRA Use only	Received By:	Date:
	CEBRA / SAC Approval:	Date:
	DA Endorsement: ( ) Yes ( ) No	Date:

Optimising New Zealand's marine biosecurity surveillance program: Review and comparison of data sources

CEBRA Project No. 170618

Milestone Report

February 2018

FINAL VERSION

Acknowledgements

This report is a product of the Centre of Excellence for Biosecurity Risk Analysis (CEBRA). In preparing this report, the authors acknowledge the financial and other support provided by the Department of Agriculture and Water Resource (DAWR), the Ministry for Primary Industries (MPI), and the University of Melbourne.

Contents

Definitions, acronyms, and abbreviations

Abbreviation / word	Definition / Description
BBWD	Biofouling and Ballast Water Declaration
CEBRA	Centre of Excellence for Biosecurity Risk Analysis
DWT	Dead Weight Tonnage
IMO	International Maritime Organization
LMIU	Lloyds Maritime Intelligence Unit
MHRSS	Marine High Risk Site Surveillance
MPI	Ministry for Primary Industries
Niche areas	Submerged surfaces on a vessel that protrude from, or are recessed into the hull, or which are not adequately protected by the antifouling coatings (e.g., rudders, propellers, stern tubes, intakes, sea-chests, internal seawater piping, bilge keels, thrusters, stabilizers, struts, grates, sacrificial anodes, dry dock support strips, etc.)
NIMS	Non-indigenous marine species
NIWA	National Institute of Water & Atmospheric Research Ltd.
TWSA	Total wetted surface area of a vessel

1 Introduction

Biosecurity New Zealand undertakes site-based surveillance of high-risk sites via the national Marine High-Risk Site Surveillance Programme (MHRSS) to detect and prevent non-indigenous marine species (NIMS) from entering and spreading in New Zealand waters. Survey methods of MHRSS are constantly being updated and improved, and there is a need to consider changes to shipping patterns and to reprioritise surveillance sites based on the relative risk of invasion and NIMS establishment at each site. Therefore, Biosecurity New Zealand aims to improve the efficiency of MHRSS by developing a systematic statistical likelihood-based methodology to assign surveillance effort to sites relative to their exposure to NIMS originating from ballast water and biofouling pathways.

As outlined in the NIWA Project Management Plan (Inglis, 2018), the objectives of this project are to:

- identify New Zealand ports with the highest relative likelihood of entry by non-indigenous marine species (NIMS) and to use this information to determine how survey effort for NIMS should be assigned among ports,
- develop a systematic, statistical likelihood-based methodology that can be used to:
  - determine the relative likelihood of NIMS entry at sites (ports and marinas),
  - select sites prior to commencing a marine surveillance programme, and
  - periodically investigate whether the Marine High Risk Site Surveillance (MHRSS) or other marine surveillance programmes are optimised for the detection of NIMS, and
- determine how any recommendations, if implemented, will affect the detection and interpretation of any long-term trends in the data set.

This project will also provide a cost-effective framework by enhancing the marine surveillance activities at the early detection level. The early detection of NIS will reduce the cost of intervention. For example, preventive measures are known to be the most cost-effective and efficient ways of minimising the impact of NIMS (Shannon et al., 2020). The Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES) estimated the cost of prevention of invasive species to be around $0.8 million a year for compliance monitoring of BW discharge in comparison to the cost of eradication which is $5 million – $20 million. Once invasive NIMS are established, the cost of loss of production, impacts on the environment, and any management costs will be $4 million - $1 billion per incursion (Arthur et al., 2015b). Data collecting programs at detection and prevention stage

can be costly and difficult because of complexity of port environments both above and below the water along with the physical, logistical, safety and legislative issues (McDonald et al., 2020). It is advisable to obtain surveillance information from a wide range of sources to help reduce the significant cost of specific surveillance activities (Arthur et al., 2015a).

Before developing the statistical likelihood-based models, exploration of all available data was required to check if they will be beneficial for the purposes of optimizing the current MHRSS. The main focus of this report is to compare the two available databases and to determine whether the project would need to purchase further data from Lloyd's (at significant cost). Similarities and discrepancies of both databases were explored to see if the data sets were materially different.

Information about vessel characteristics and vessel traffics were available from two databases, one provided by the Ministry for Primary Industries' Intelligence and Targeting Team (ITT) and the other from Lloyd's Maritime Intelligence Unit (LMIU). Vessel arrival data contained in the MPI database were compared with the Lloyd's commercial vessel database that was recently acquired by MPI for accuracy and specificity of the data. If there were no substantial difference between the two data sets, then a combination of data provided by MPI, NIWA, and already purchased from Lloyd's can be used for modelling to inform high risk sites for NIMS arrival and establishment through ballast water and biofouling pathways. The approach used to develop the statistical likelihood-based methodology assumes that the likelihood of successful establishment of NIMS within a site is related to the number of species ('colonization pressure') and the total number of individuals ('propagule pressure') that the site is exposed to (Lockwood et al., 2005; Lonsdale, 1999). The total biofouling mass and the volume of ballast water discharged per port are predicted assuming that the total mass of the risk commodity transported in these pathways are related to propagule pressure and colonization pressure.

The comparison of MPI and Lloyd's data sets are done using the overlapping data available for this study, but more recent data will be used to make predictions based on the likelihood-based model. Several years of data from 1998 to 2008 (named *historical* data) were available for model building to predict the ballast discharge volume as a function of voyage properties and vessel characteristics, e.g., arrival port, vessel type, ballast capacity or DWT, and the reported 'intent to discharge'. The models will be used to predict average annual port-level ballast water discharge per voyage using more recent 2015-2017 data (named *contemporary* data). Similarly, the historical data will be used to fit a statistical model to predict the total annual port-level biofouling mass on vessels arriving in New Zealand between 2015 and 2017. The vessel characteristics (e.g., age of antifouling, period of

inactivity, vessel type, size, and speed) and voyage features (e.g., first port of arrival) are important predictors in model building for biofouling mass. The first port of arrival and the other ports that the vessel visits during each journey are important in ballast water analysis because the vessel might discharge at several ports within a journey and only visit other ports. For instance, a vessel that visits Picton, Tauranga, Auckland, and Christchurch might only discharge ballast water at the third and fourth port. The first port of arrival is a key variable in biofouling analysis and most likely to receive propagules but not all the visiting ports because biofouling is assumed to be consistent or slightly different at the ports a vessel visits next. Thus, in this report, a vessel's signalled intent (MPI) and the commercial (Lloyd's) databases were compared using a method that considered the vessel movement. The analysis can also inform what, if any, further data could be collected by MPI in the future to continuously update and inform the models of high-risk sites that are to be developed in the next steps of this project.

Another challenge faced by this study was to extract ballast water discharge information from PDF files provided in the New Zealand Biofouling and Ballast Water Declaration forms (Appendix A – section 6.1). These were PDF forms that were completed by vessel captains prior to arrival in NZ as to their intention to discharge ballast water in NZ territorial waters. Due to the format of these documents, it was difficult to extract the data, therefore another objective of this report was to investigate a process to automate the extraction of ballast water discharge data from PDF forms.

This report is structured as follows. Section 2 starts by describing the two data sets and the sources of data. It then explains the steps taken to transform the port names into strings for each individual journey. This is followed by a Levenshtein distance analysis to measure the dissimilarity between the strings, and a generalised linear mixed model (GLMM) to determine the variables responsible for these discrepancies. The rest of this section explains how the data were extracted from the PDF files. Sections 3 and 4 are assigned to the results, discussion, and recommendations for improvement of the data acquisition and curation. More information about the data sets, data cleaning, data preparation, more visual comparison of data sets, and other information are provided in section 5 (Appendix A).

2 Methods

2.1 Sources of data

The MPI data used in this report was provided by the Ministry for Primary Industries' Intelligence and Targeting Team (ITT), and the Lloyd's data was collected by the Lloyd's Maritime Intelligence Unit (LMIU) and purchased by MPI. These data sets contained information about vessel characteristics and shipping traffic of the international vessels that arrived in New Zealand ports during years 2000–2005 and 2016 for Lloyd's data and from 2012 to 2017 for MPI data. The records related to the vessel characteristics and vessel traffic including vessel type, vessel movement and journeys in two data sets were matched using their unique International Maritime Organization (IMO) numbers. IMO number is a unique vessel identifier and remains invariant to changes to other vessel features, such as the name and flag of the vessel. The data description, cleaning and preparation steps were described in sections 6.1 and 6.2 of Appendix A in more details. The exploratory data analysis was done using the whole data sets, but the remainder of the analysis focuses only on year 2016 to remove any source of discrepancy between the data sets. The initial analysis assessed the similarities and differences in the MPI and Lloyd's datasets. The data comparison and matching the records of year 2016 are presented in section 6.4 of Appendix A. After the data cleaning and merging, 920 vessels with unique IMO number remained in both data sets with 6871 and 6226 port visits for common vessels for Lloyd's and MPI data, respectively. All the selected variable used in this report were 'IMO number', 'year', 'vessel type', 'generic type code', 'grouped vessel type', 'movement sequence', 'flag', 'last country', 'first port', and the 'visited ports'. The description of these variables and their categories were presented in Appendix A.

2.2 Data analysis

2.2.1 Levenshtein distance (LD)

Levenshtein distance (LD) measures the similarity between two strings by counting the number of insertions, deletions or substitutions required to transform one string into another (Levenshtein, 1966). As no transformation is required when two strings (named source and target strings) are identical, lower edit distance values imply less difference and greater values indicate more difference between two strings (Lazreg et al., 2020). This analysis was used to determine the port-to-port differences recorded for each individual vessel in the Lloyd's data compared with MPI data.

Figure 1 demonstrates the steps taken to transform the port names into strings for each individual vessel journey in each of the datasets. By converting each port into a letter (stored as variable 'port string'), a string was generated for each individual vessel journey in each of the datasets (Figure 1). After assigning a letter to each port (step 1), a code (variable 'movement ID') was built for each journey using the unique IMO number and report number (step 2). A journey starts when a vessel arrives in NZ and ends when it travels to another international port, and vessels can have multiple journeys to NZ in a single year. For example, a vessel with IMO number of '8067880' has travelled to New Zealand four times in 2016, with each journey identified by a different report number. This vessel has visited Wellington, Nelson, Auckland, and again Wellington in April, June, July, and October in the same year, with 'movement number' of 1, 2, 3, 4 and 'movement ID' of a, b, c, d for each journey. For each combination of unique IMO number and 'movement ID', all the visited ports (letters from step 1) at each journey are placed together to form a string for each data set (step 3). As it can be seen from Figure 1 – step3, the same vessel with IMO number of 8067880 only visited one port at each journey in MPI data but visited several ports in Lloyd's data. For example, for the journey with 'IMO-movement ID' of '8067880 a', the 'MPI string was 'm' and the 'Lloyd's string' was 'aagaa'. This means that Wellington was the only recorded port in MPI data for that journey whereas Auckland, Auckland, Nelson, Auckland, and Auckland were recorded as visiting ports in Lloyd's data. These generated port strings were used to compare the records for individual vessels between the datasets. LD analysis was done using function '*stringdist()'* from the stringdist package in R (van der Loo, 2014).

Step 1

Port	Alphabet
Auckland	a
Bluff	b
Dunedin	c
Gisborne	d
Lyttelton	e
Napier	f
Nelson	g
New Plymouth	h
Picton	i
Taharoa	j
Tauranga	k
Timaru	l
Wellington	m
Whangarei	n

Step 3

No.	IMO-movement ID	Lloyd's string	MPI string
1	7416480 a	llbllmml	l
2	8000880 a	mbgmelllllllmmmem	m
3	8000880 b	m	g
4	8067880 a	aagaa	m
5	8067880 b	flfa	g
6	8067880 c	a	a
7	8067880 d	a	m

Step 2

No.	IMO number	Report number	Visiting port	date	Movement	
					number	ID
1	7416480	AYG3457821	Timaru	2016-09-12	1	a
2	8000880	ANY3823821	Wellington	2016-05-18	1	a
3	8000880	ANY3823430	Nelson	2016-10-15	2	b
4	8067880	ANC3215532	Wellington	2016-04-13	1	a
5	8067880	ANC3215827	Nelson	2016-06-08	2	b
6	8067880	ANC3215902	Auckland	2016-07-08	3	c
7	8067880	ANC3216546	Wellington	2016-10-22	4	d

Figure 1. The steps taken to generate a string from the visiting ports for each individual vessel journey in each of the datasets.

2.2.2 Generalised Linear Mixed Modelling of Levenshtein distance

Using the LD for each vessel calculated from the individual journey string and ordered alphabetically, we conducted generalised linear mixed models to determine the most important predictor variables to explain the differences in the MPI and Lloyd's data sets (that is factors that led to higher string distances). 'IMO number', 'flag', 'first port', 'last country', 'grouped vessel type' and 'visited ports' were the variables used in the GLMM analysis.

2.2.3 Extracting data from PDF files

Almost 500 Biofouling and Ballast Water Declaration forms were available in PDF format, which were submitted by international vessels arriving to NZ as part of their biosecurity obligations. The forms provided details on the number of ballast tanks and tank capacity of the vessel, whether the vessel intends to discharge any ballast water in NZ and if so, the ballast water management method to be undertaken, volumes to be exchanged, and original source of the ballast water. These forms were a signal of intent. These were the only ballast data available for international vessels arriving in NZ which could be used for modelling to inform high risk sites for NIMS arrival and establishment. Due to the large numbers of PDF forms, we investigated automating the extraction of the relevant data. The

data required from each form were the vessel name, IMO number, the first port of arrival, ballast water sources and volumes of approved tanks.

Converting PDF to text

Converting PDF files into a character vector in R was straightforward using the 'pdftools' package (Ooms, 2017). The '*readPDF*' function could also be used in the 'tm' package in R as an alternative which allows writing customised functions. Using the 'strsplit' tool, it was possible to split the elements of the character vector into substrings based on '\r', which indicates a new line and is added when converting from PDF to character vector. Another option was to batch convert all the PDF files into Comma Separated Values (CSV) files which could then be read into R. This was done using the Action Wizard in Adobe. The steps were 1) Create new action; 2) select Save and Export; 3) select Save (which opens a table); 4) then click on "specify settings"; 5) export files to alternate format and select excel workbook. The batch of files were then saved as CSV files and imported into R using '*read.csv*' function.

Extracting text

The package 'stringr' (Wickham, 2019) can be used for character manipulation, removing white space, and pattern matching with regex. Using the function '*str_extract*' and regex code, the PDF content, e.g., Vessel Name, IMO number, Arrival Port, and Arrival Date, Number of tanks in ballast, and Volume could be extracted. These steps were encoded to automate data extraction from each form and put the data into a data frame. Question 6 in Part 1 was a checkbox that asked: "are any ballast tanks intended for discharge, or possible discharge, in NZ ports or territorial waters"– if this was marked 'Yes,' then vessel owners were required to complete Part 3 of the table. We were unable to determine how to extract data from PDF check boxes in R. To get around this, we assumed that if there were data provided in Part 3, then the response to question 6 would be 'Yes' and if there were no data, then the response would be 'No'.

3 Results

The visual inspection of the vessel traffic for Lloyd's (years 2000–2005 and 2016) and MPI (2012–2016) data sets showed a general rise in the number of vessels arriving in New Zealand in the last decade, especially at ports such as Tauranga, Wellington, and Whangarei (Figure 1 in Appendix A). Ports Tauranga, Auckland, and Lyttleton had the highest number of visits in both data sets. Tauranga had the largest number of vessel arrivals followed closely by Auckland in both data, except that Auckland had the greatest number of arrivals for the

Lloyd's 2016 data. The visual comparison of the data showed a rapid elevation in vessel traffic in 2016 for the two databases, MPI and Lloyd's, accompanied by a rise in the number of bulk/oil vessels and a fall in the number of cargo carriers. In both data sets, containers had the largest arrivals in New Zealand. For most of the years, cargo vessels had the second highest number of visits in Lloyd's, whereas bulk/oil vessels were the second most frequent vessel types in MPI data. The changes in vessel traffic of different vessel types arriving in New Zealand's ports were compared in more details in Appendix A – section 6.3.

As explained in data preparation sections, the main analysis of LD and GLMM was conducted using 2016 data. The relationship between the port strings of two data sets, prepared for LD analysis was explored by plotting the total string lengths in a scatterplot (Figure 2a). This figure provides an indication of string differences for the total number of port visits for individual vessels during 2016 in MPI and Lloyd's databases. More than 90% of the strings in both datasets were shorter than 20 characters. Almost 9% of MPI strings were longer than Lloyd's strings whereas 29% of Lloyd's strings were longer than MPI strings. For example, as Figure 2a shows, for the string length of 1 in MPI data, there were multiple strings with lengths greater than 1 in Lloyd's data. This is more obvious in Figure 2b that presents total string lengths in MPI and Lloyd's data sets on a log10-transformed scale.

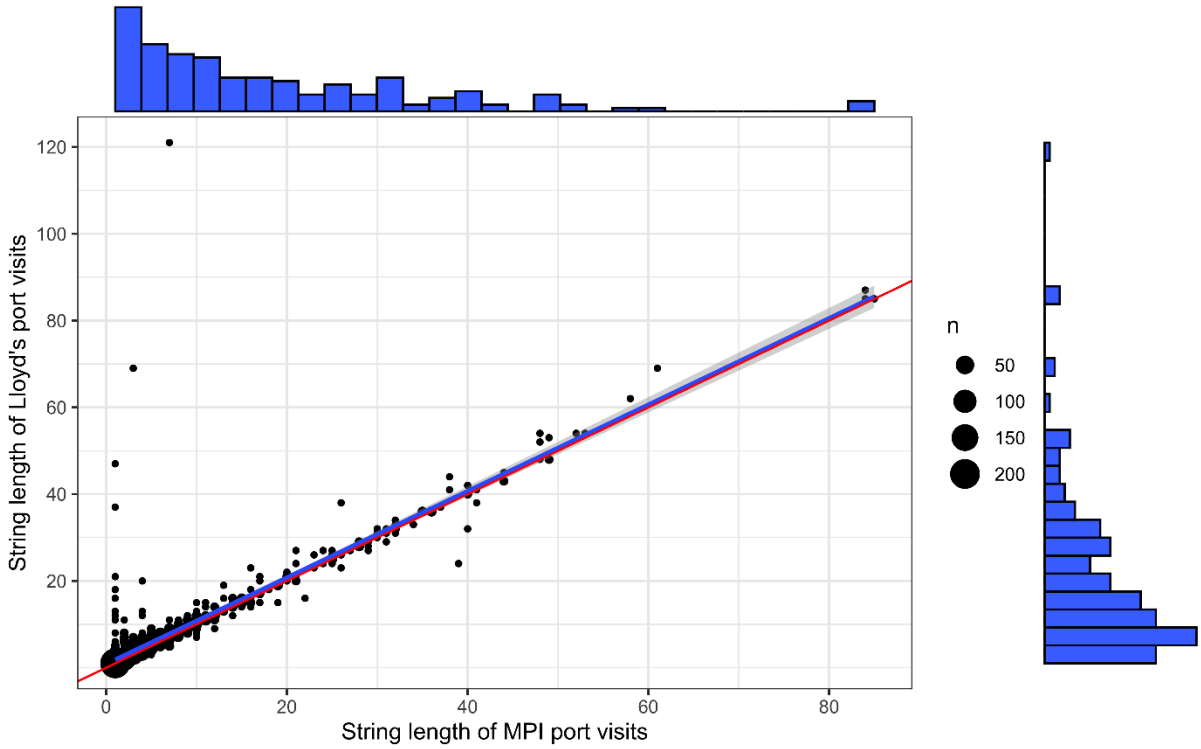

Figure 2a. Scatterplot of total string lengths calculated based on the number of port visits for each vessel in MPI and Lloyd's data sets in 2016. A regression line in blue colour and a 1:1 line in red colour were overlaid the points in the graph.

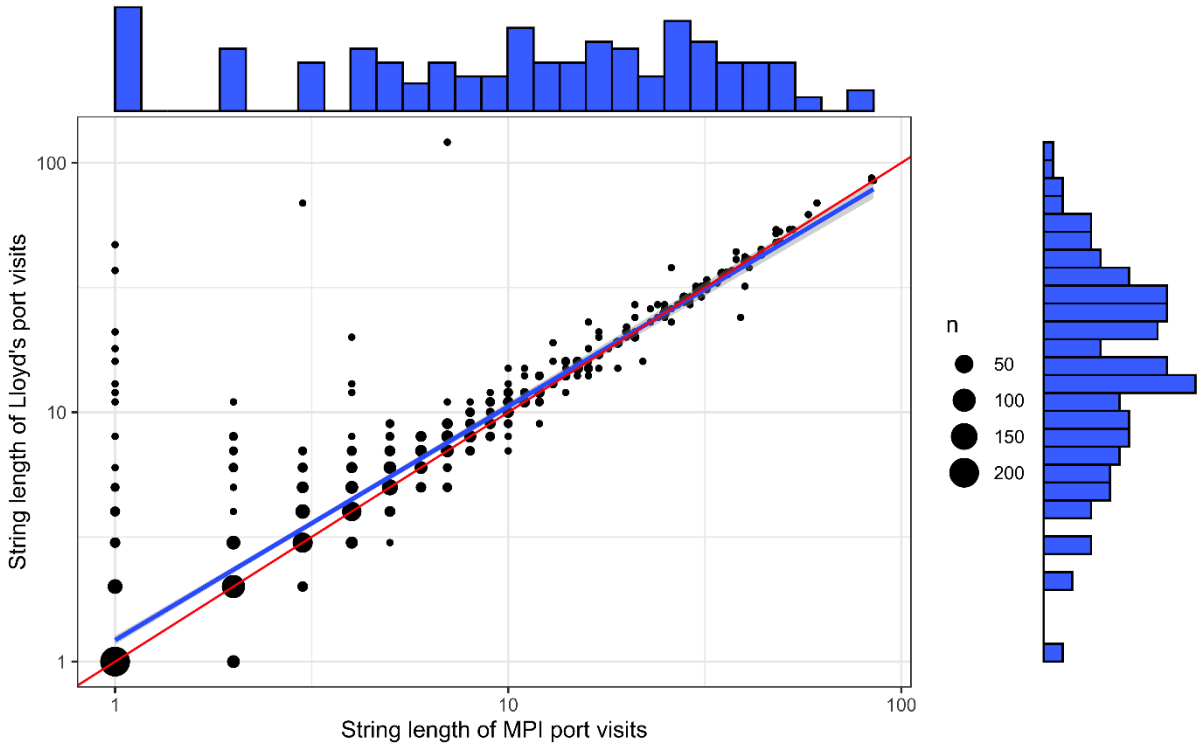

Figure 2b. Scatterplot of total string lengths calculated based on the number of port visits for each vessel in MPI and Lloyd's data sets in 2016 (on a log10-transformed scale). A regression line in blue colour and a 1:1 line in red colour were overlaid the points in the graph.

The Levenshtein distance was calculated for four different scenarios; single string for all journeys of each vessel, reordered single string alphabetically (Figure 3), string for the individual journey of each vessel, and reordered journey string alphabetically (Figure 4). The statistics, e.g., median, mean and maximum calculated for these four cases are presented in Table 1. Figure 3 illustrates the values of Levenshtein distance when a single string for each vessel was used to compare both data sets. That is, the total port visits by a vessel in 2016 were compared irrespective of whether they were considered different journeys. In this graph, the LD related to the port visits recorded in the order visited by each vessel (group A in red colour), were overlayed by the port visit strings reordered alphabetically (group B in blue colour). When the port visit strings were reordered alphabetically, to assess whether the identity of visited ports was the same (regardless of order), the median of LD decreased from 1 to 0 and the mean decreased from 2.26 to 1.19. The number of vessels with an LD greater than 10 was 113 in the former which decreased to 31 in the latter (Table 1). The 920 vessels made 2433 journeys in each dataset, but theses journeys were not exact matches. There were 92 discrepancies between Lloyd's and MPI journeys, i.e., Lloyd's had 46 journeys not in MPI and MPI had 46 journeys not in Lloyd's. These journeys were removed while building the graphs. When the individual journeys of each vessel were considered in the LD analysis (Figure 4 – group A), the LD median remained as 0 but the mean decreased to 1.1 and the maximum LD value decrease from 114 to 78. In the individual journey case, there were only 9 vessels with an LD greater than 10 in comparison with 35 vessels in port strings with all journeys combined (in single string scenario). (Figure 4, Table 1). When the port visit strings were reordered alphabetically for each vessel (Figure 4 – group B) the results were much more similar, and the plots in this figure almost overlapped. The LD mean in ordered strings was 1.01 for each vessel which was slightly lower than the strings without order with the mean of 1.1 (Table 1). The results of pairwise comparison of LD values using t-test for these four different scenarios showed that the differences were significant (Table 2).

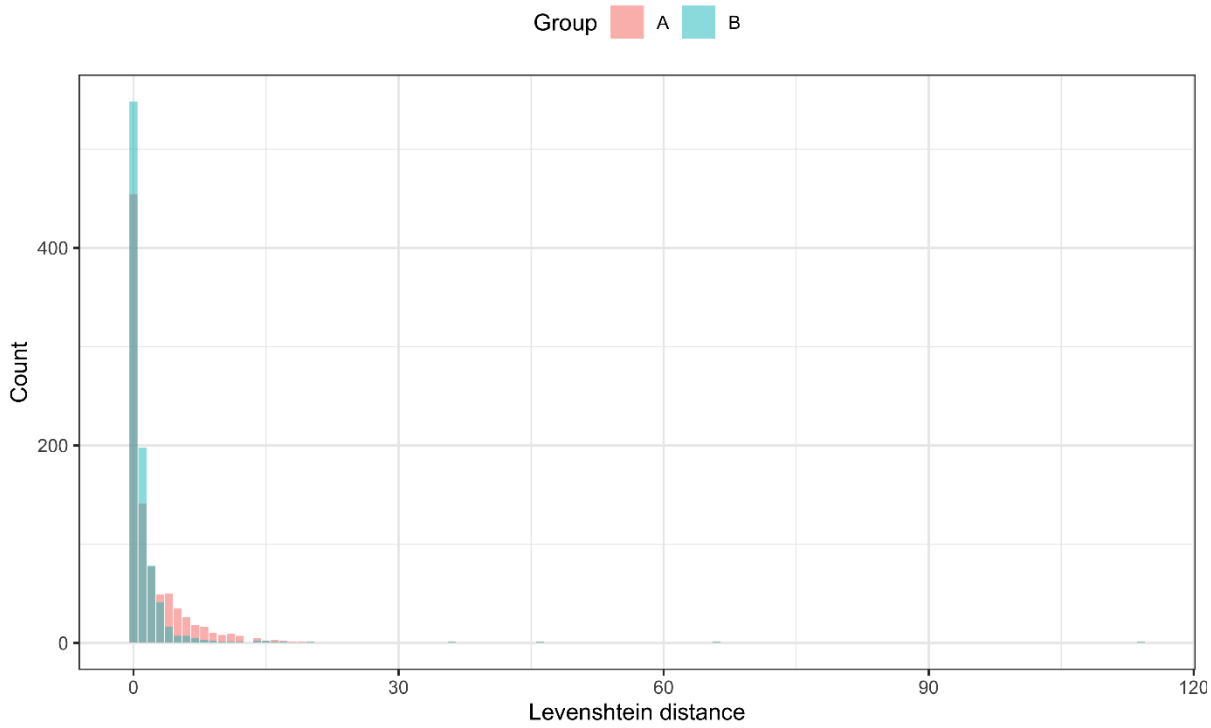

Figure 3. Levenshtein distance of MPI and Lloyd's data for all port visits for each vessel, i.e., calculated using a single string for each vessel (n = 920). In group A illustrated in red colour, the port visits were recorded in the order visited by each vessel whereas in group B represented in blue colour, the port visit strings were reordered alphabetically.

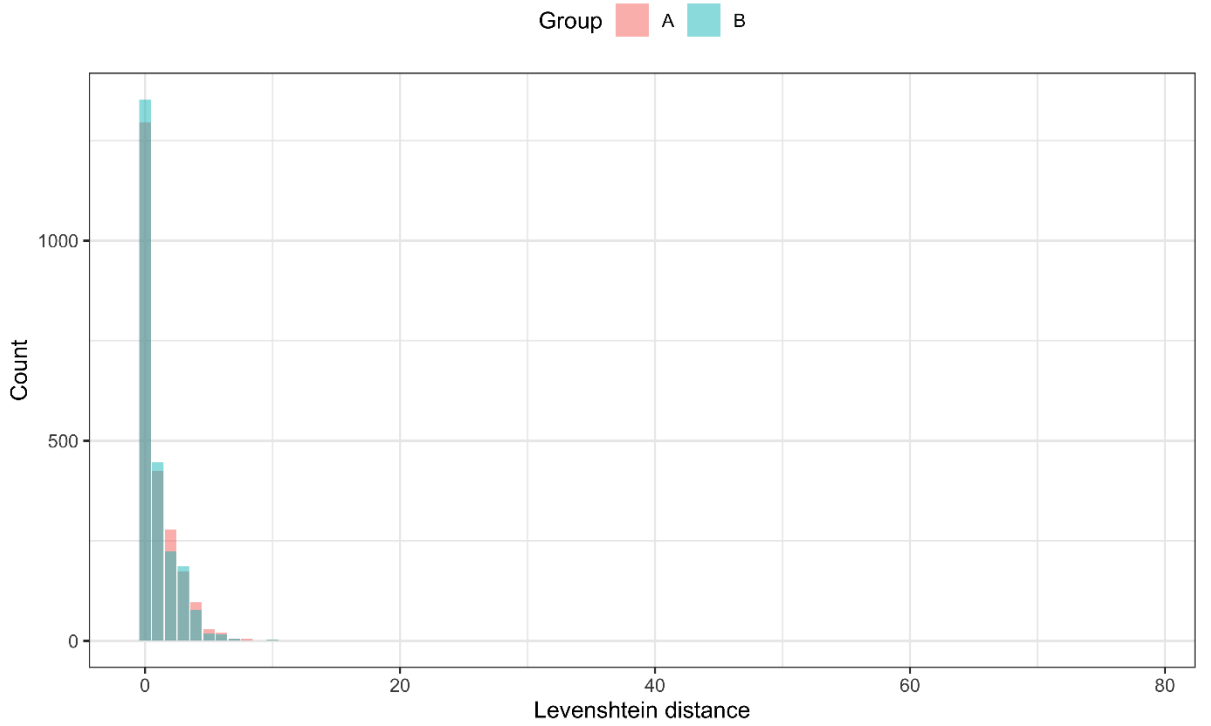

Figure 4. Levenshtein distance of MPI and Lloyd's data for all vessel journeys, i.e., calculated using a string for each journey per vessel (n = 2433, NA = 92). In group A illustrated in red colour, the port visits were recorded in the order visited by each vessel whereas in group B represented in blue colour, the port visit strings were reordered alphabetically.

Table 1. The statistics calculated for a single string for all journeys of each vessel, reordered single string alphabetically, a string for the individual journey of each vessel, and reordered journey string alphabetically.

Levenshtein distance statistics	String for all journeys per vessel		String for an individual journey per vessel	
	Not ordered	Ordered	Not ordered	Ordered
Median	1	0	0	0
Mean	2.26	1.19	1.1	1.01
Maximium	114	114	78	78
LD > 5	113	31	43	37
LD > 10	35	13	9	9
LD > 20	4	4	4	4

Table 2. The results of t-test comparing Levenshtein distance values calculated under four scenarios of single string for all journeys of each vessel (LD1), reordered single string alphabetically (LD2), string for the individual journey of each vessel (LD3), and reordered journey string alphabetically (LD4). t-value, df, mean of the differences (confidence interval) for each pairwise comparison are given.

Variables	t-value	df	Mean (CI)	P-value
LD1, LD2	3.43	2356	0.12 (0.05, 0.18)	0.001
LD1, LD3	-41.11	2356	-3.4 (-3.56, -0.32)	< 0.001
LD2, LD4	-6.77	2356	-0.4 (-0.5, -0.29)	< 0.001
LD3, LD4	42.31	2356	3.11(2.97, 3.25)	< 0.001

To determine whether there were any ports with visit discrepancies between the MPI and Lloyd's data, we calculated the number of port visits by all vessels with a LD score greater than zero. These data are provided in Table 3. Tauranga had the greatest number of differences, and by a substantial margin, followed by Whangarei and Auckland. Tauranga and Auckland followed by Lyttelton and Napier had the highest proportion of visit discrepancies among sites. The proportion of visit discrepancies per port followed the same trend in both data sets. For example, almost 20-23% of visits in Tauranga, 17-18% of visits in Auckland, and 10% of visits in Lyttelton had LD score greater than zero in both MPI and Lloyd's datasets.

Table 3. The number of visits per port for all vessels with a Levenshtein distance greater than 0. The total number of visits at each port are given in the parenthesis. The proportion of visits per port in each data set and absolute differences of number of port visits between two data sets are also provided.

Ports	Number of visits for LD>0		Proportion of visits (%)		Differences between data sets relative to MPI visits (%)
	Lloyd's	MPI	Lloyd's	MPI	
Auckland	972 (1291)	885 (1204)	17.62	18.57	10
Bluff	178 (230)	163 (215)	3.23	3.42	9
Dunedin	326 (392)	320 (386)	5.91	6.71	2
Gisborne	105 (133)	101 (129)	1.90	2.12	4
Lyttelton	540 (691)	475 (626)	9.79	9.96	14
Napier	493 (616)	458 (581)	8.94	9.61	8
Nelson	284 (318)	258 (292)	5.15	5.41	10
New Plymouth	255 (304)	189 (238)	4.62	3.96	35
Picton	58 (76)	53 (72)	1.05	1.11	9
Taharoa	7 (21)	8 (22)	0.13	0.17	13
Tauranga	1279 (1567)	998 (1286)	23.18	20.94	28
Timaru	255 (282)	224 (251)	4.62	4.70	14
Wellington	455 (535)	427 (509)	8.25	8.96	6
Whangarei	310 (415)	208 (313)	5.62	4.36	33

The generalised linear mixed model, as written in Equation 1, indicated that vessel 'flag', the last country, and the 'grouped vessel type' were indicators of having a high LD score. IMO number was included in the model as a random effect to consider. The model containing these variables was significantly different (χ^2 (111) = 347.5, p < 0.001) from the null model including only intercept (ΔAIC = 135).

$$Levenshtein\ Distance = Normal(\mu_{is}, \sigma^2)$$

$$\mu_i = \alpha_1 + \beta_1 \times flag_{is} + \beta_2 \times last\ country_{is} + \beta_3 \times grouped\ vessel\ type_{is} + \varsigma_i$$

Equation 1

In Equation 1, ς_i is a random intercept with mean 0 and variance σ^2. Results of ANOVA test showed that 'flag' (Wald $\chi^2(45) = 292.34$, $p < 0.001$) and 'grouped vessel type' (Wald $\chi^2(8) = 107.49$, $p < 0.001$) were significant, but 'last country' (Wald $\chi^2(58) = 46.30$, $p = 0.86$) was left in the model because the model had a smaller AIC while including this variable in the model. The estimated parameters from the GLMM model and the model comparison using AIC are given in Appendix A – section 6.5.

4 Discussion and recommendations

Inconsistencies between the MPI and Lloyd's data sets were explored prior to development of statistical likelihood-based modelling technique to reprioritize surveillance

sites aligned with likelihood of NIS entrance at each port. Based on the exploratory data analysis, the number of vessel traffic in both data sets indicated an increase in the number of vessels arriving in New Zealand in the last decade. In both datasets, Tauranga, Auckland, and Lyttelton were the busiest ports and experienced a higher number of visits in 2016, except for Lyttelton that received less visits in Lloyd's data in the same year. According to Lloyd's, between 2005–2016, there was an increase in vessel arrivals of more than 30% in New Zealand, equating to almost 2,500 additional port visits. This increase was accompanied by a change in vessel categorisation, e.g., the number of container and bulk/oil vessels increased while cargo vessels decreased in 2016. In general, both data sets showed similarities in the number of visits by vessel types, there were differences in the port visits by each vessel type. For example, similar to as seen the Lloyd's data, containers had the highest number of visits in MPI data and mostly arrived in large ports. These vessels mostly visited Auckland, Tauranga, Wellington, Nelson, Lyttelton, and Napier in Lloyd's data whereas Auckland, Tauranga, Lyttelton, Napier, Dunedin, and Wellington were the visiting ports by this vessel type in MPI data.

Despite a rather similar trend in changes in the number of visiting ports in both datasets, there were slight differences in the number of port visits for each vessel. Generally, there was an agreement between both datasets in terms of the length of strings generated from port names, indicating a rather similar number of port visits by each vessel. However, there were vessels in each dataset with a higher number of visiting ports than the other dataset, especially this was the case for Lloyd's data with longer stings. Most of these long strings comprised several journeys per vessel, but not all of them. The measured edit distance scores had zero mean and low median implying high similarity between port visits for each individual vessel in the Lloyd's compared with MPI data. These scores were reduced significantly when the total port visits by a vessel were compared considering different journeys. After taking different journeys per each vessel into account, there were still individual journeys with different visiting ports in Lloyd's compared with MPI data. For example, an individual journey for a vessel might be a single visit to Auckland in MPI data, but multiple visits to other ports in Lloyd's data. The order of the visiting ports by each vessel was also a source of discrepancy between the two data sets. The dissimilarity edit distance was considerably lower when the port visit strings were reordered alphabetically for each vessel, i.e., if the order of visiting ports in each journey did not matter.

All the ports showed visit discrepancies between the MPI and Lloyd's data as they had visits with a LD score greater than zero and these differences varied systemically between ports. These discrepancies were proportional to the number of visits they received. The

highest proportion of visit discrepancies belonged to Tauranga, Auckland, and Lyttelton with the highest number of port visits in both data sets. According to GLMM results, 'flag', 'the last country', and the 'grouped vessel type' were most important predictor variables that explained the differences in the MPI and Lloyd's data sets measured by Levenshtein distance. Vessel type was related to the discrepancy probably because each the datasets had a different categorisation system for vessel types. That is, the data sets were not consistent in assigning vessels in the same category. For example, there were 12 vessel types in Lloyd's date versus 10 vessel types in MPI data, with research and dredge vessel type missing from the latter. Although the broader category of 'grouped vessel type' with eight categories shared in both datasets was used, the source of discrepancy persisted: especially because the categories used in the analysis were from MPI data merged into the dataframe containing LD scores for MPI and Lloyd's data. Another reason was inconsistency between the datasets in recording the visiting ports for each vessel type. For instance, vessel type tug had the highest recorded number of visits in New Plymouth in Lloyd's data whereas vessel types of container and bulk/oil were most frequent in this port in MPI data in 2016. Other variables related to LD scores were 'flag' and 'the last country', so there seem to be differences between the datasets in terms of the flag and the last country recorded for some of the vessels. In another work conducted by Institute Superiore Mario Boella in Italy to improve automatic recognition of port names transmitted by vessels, misspelling the port name, port code, and country name was a source of discrepancy between database and incoming data. They used Levenshtein Distance to determine destination and source ports while matching the strings of transmitted port names or codes with the ports details in database (Morisio et al., 2018). The data used in this report were gathered in written formats, e.g., filling offline electronic PDF forms and sending them back via email, and the extraction of information from these files are prone to errors. Current process of data acquisition which relies on offline forms has limited mechanism for data validation. This increases the likelihood of errors in the supplied data in both stages of entry by the vessel Master and when the information is recoded to a secondary data management system.

In this study, a process was investigated to automate the extraction of ballast water discharge data from PDF forms. To do so, using several packages in R, the PDF files were transformed to text and the information was extracted from the texts, subsequently. There were issues facing this process handling check boxes or extraction off all the required information from the text regarding the volume of discharge. Due to difficulties in data extraction involving check boxes, the assumption had to be made based on the information in other parts, especially for the yes/no answers to the question related to the intention of ballast

water discharge. The reliability of the answers to this question were checked by cross referencing these answers with the volume of ballast water reported to be discharged by each vessel. As there was not good agreement between the answers to the intention of discharge and the reported discharge, this predictor was not used in the main analysis.

Based on the findings of this study, the MPI 2016 data were not sufficiently different to justify the expense of purchasing the Lloyd's data (at a substantial cost). If the results had indicated a considerable difference, then a value of information analysis (VoI) could be useful to decide about the cost incurred to MPI in return for the benefit from more data, but we consider this beyond the scope of the current report. VoI analysis is useful to weigh the costs and benefits of different monitoring and research options for removal of uncertainty (Bolam et al., 2019; Heath et al., 2016). For example, if Lloyd's data was proved to contain more information than MPI data about vessels that bring higher risk to New Zealand ports, it would be worth considering the cost of purchasing that data for the benefit of reducing the uncertainty related to that. This would be possible by comparing the expected performance of the surveillance designs that have been generated using models developed by the two different sets of data. An overview of the utility of the MPI and Lloyd's databases for quantifying port-to-port traffic in this report did not justify purchasing more data.

Lloyd's recorded many New Zealand domestic journeys of vessels such as the Cook Strait ferries which needed to be removed for analysis. The MPI database contains data on vessels less than 100 tonnes, including various types of international yachts. These data would be valuable for risk modelling for high-risk sites in New Zealand. The Lloyd's data also lacked the port sensitivity that MPI data had, for example, Lloyd's in some cases combined Opua into Auckland, and Akaroa into Christchurch. When cross referencing some of the cruise liners that berth in Opua from MPI data, they were recorded in Lloyd's as arriving in Auckland. Opua and Akaroa are both currently considered high-risk sites and data on the vessels arriving in each of these ports would be valuable for risk modelling. Lloyd's 2000 – 2005 had data for 27 NZ ports, including Opua, but only 9 visits across the 6 years of the data. It would be worth enquiring whether Opua (as an example) had on occasion been combined into Auckland in these data as it hah for 2016, and whether Lloyd's had aggregated some of the ports between their earlier and later data. There were journeys that were recorded in MPI data but not in Lloyd's data and vice versa. MPI data were provided by vessel captains on arrival into NZ of their intended domestic travel. This intended travel may not have occurred, which may explain some of the discrepancies between the Lloyd's and MPI port visits. MPI data had the advantage of only recording international arrivals and their intended destinations within NZ, thereby automatically excluding these types of trips from the database. The major

downfall of this, however, was that a vessel may not follow through with its signal of intention, leading to false reporting in the database. This comparison work was partially useful to quantifying these discrepancies. The finding of this study can be beneficial in updating current guidelines on data acquisition and curation while gathering data by MPI for continuously updating and informing the models of high-risk sites, developed in the next steps of this project. A dataset comprised of MPI data (and data already purchased from Lloyd's by MPI) will support model building for biofouling mass and BW discharge. The vessel characteristics and voyage features such as vessel type, arrival port, and dead weight tonnage from historical data (1998 - 2008) will be used to predict discharge port, discharge volume, and biofouling mass for contemporary data (2015 - 2017). Then, entry likelihood scores for each port will be calculated from ballast water discharge and biofouling exposure which will subsequently be utilised to allocate surveillance effort among sites.

4.1 Recommendations for data acquisition

As a significant proportion of the resource available to this study was consumed in data acquisition and curation in preparation of the main analysis, a few recommendations are made here so that marine biosecurity risk profiling information can be used in future strategic purposes. Data collection using offline forms has limited mechanism to validate data and is susceptible to errors during entry or re-coding into management systems. Shifting to online forms accompanied by a vessel check system, similar to what USA and Australia have recently adopted, will improve the efficiency of data collection and validation.

The amount of free text in the spreadsheet currently used by ITT Target Evaluators (the 'MPI Craft Work Schedule') could be restricted and replaced by standardised answers for specific columns (i.e., drop-down lists) that apply to all vessel types. It is also suggested to set up forms which allow for the straightforward extraction of data, and we recommended that this is investigated and deployed. In its current form, these data would require substantial resources to extract into a usable format for analysis. The forms formatting is suggested to be improved so that tables, dates, and other information within each form are more consistent.

Finally, the proposed volume of ballast water discharge is not required to be reported in the BBWD in its current form. This was required in New Zealand until 2017 and the reporting form in BW management guidelines had assigned a section for this purpose (Marine Environmental Protection Committee, 2018). In Australia and USA, such information on ballast water discharge is considered a requirement on the ballast water reporting forms. Gathering data on BW volume is recommended because it will help ballast water management and compliance auditing to be evaluated in a strategic and more efficient way.

5 Appendix A

5.1 Data overview

Several data sets were provided by MPI, NIWA, and Lloyd's for this study and are summarised in Table 1 derived from Inglis (2018). The first three rows of this table contain the data used to build models to optimise the New Zealand's marine biosecurity surveillance programme (Hatami et al., 2021). The marine surveillance data (3rd row in Table 1) provided by MPI contained 36235 records of vessel arrivals by New Zealand ports during 2012 – 2017 and is called MPI data in this study. The data provided by Lloyd's / NIWA from domestic vessel movements study contained 43592 records of vessel arrivals by New Zealand ports during 2000 – 2005 (5th row in Table 1). The data provided by Lloyd's contained 9616 records of vessel arrivals by New Zealand ports in 2016 (6th row in Table 1). These two last data sets form the Lloyd's data in this study. The data that used for Levenshtein analysis were extracted from the datasets highlighted in grey in the table. More information about these datasets and details of data preparation are documented in part 2 and part 3 of Appendix A – sections 6.1 and 6.2.

Table 1. Summary of data available for developing a risk model for entry of NIMS to New Zealand shipping ports, taken from NIWA Project Management Plan (Inglis, 2018).

ID	Description	Source	Period	No. records	Potential use	additional information
1	New Zealand Biofouling and Ballast water Declarations (Historic)	MPI	1998 - 2008	15745	Use to parameterize a predictive model of ballast discharge.	This data set contains complete data from biofouling and Ballast Water Declarations but requires grooming to allow analysis.
2	Vessel Biofouling Characterization study (Historic)	MPI /NIWA	2004 - 2007	508	Use to parameterize a predictive model of biofouling on arriving vessels.	This dataset contains measures of biofouling on international vessels arriving in New Zealand ports and information about maintenance and voyage history of arriving vessels.
3	Information report: Marine Surveillance data	MPI	2012 - 2017	36235	Predicting biofouling and ballast water risk.	This dataset contains vessel arrivals by New Zealand ports during 2012 – 2017, but the joined summary fields cover only the period 2015 – 2017.
4	New Zealand Biofouling and Ballast Water Declarations	MPI	2016	998	Potentially use this subset of records to determine the biofouling and ballast water risk for each individual vessel	- Keyed in data for all questions in the Biofouling and Ballast Water Declaration - Only 473 of 998 declarations have been keyed into an electronic format.
5	Domestic Vessel Movements Study (Historic)	Lloyd's / NIWA	2000 - 2005	43592	Compare historic risk profile of NZ ports with current risk profile.	- Vessel arrivals by New Zealand port (2000–2005)
6	Vessel arrivals to NZ Ports	Lloyd's	2016	9616	Use vessels tables to calculate TWSA and the area of niches for each arriving vessel.	- Vessel arrivals by New Zealand ports in 2016

5.2 Data preparation

5.2.1 Lloyd's data set

Data purchased by MPI from the Lloyd's were supplied to this study as several separate files for years 2000 – 2005 and 2016 that required cleaning and preparation before the analysis. Data for years 2000 – 2005 were built by combining several excel files containing information about the vessel characteristics ('*Ld_vessels_2000-2005.csv'*), vessel movements (*'Client_place_moves2000-2005.csv'*), the places the vessels visited ('*Ld_places_2000-2005.csv'*), and vessel types ('*Ld_vessel_types.csv*'). In the file related to the vessel types, there were 14 levels of generic type codes including B, C, D, F, G, L, M, O, P, R, T, U, X, and Y that represented bulk, bulk/oil, dredge, fishing, cargo, LNG/LPG (Liquefied Petroleum Gas Carrier), VPL (vehicle/livestock carrier), other, pass/roro (passenger), research, tanker, container, tug, and drill, respectively. The files containing information about the vessel characteristics and vessel types were merged based the unique generic type codes. Files related to vessel characteristics and vessel movements were merged by shared LMIU number, and files related to vessel movements and the visited places were merged by shared place ID number. Both files resulted from this merging were combined to build a master file with 43822 records with 2499 unique LMIU number. To simplify the comparisons between the data sets, some of the places were removed or renamed; for example, Mount Maunganui was renamed to Tauranga, and Port Chalmers was renamed to Dunedin. The visits to Chatham Islands, Doubtful Sound, Greymouth, Milford Sound, New Zealand, Opua, Stewart Island, Tarakohe, Westport, and Whakaaropai Terminal were removed from the data. The final Lloyd's master file for years 2000 – 2005 consisted of 43147 records with 2472 unique LMIU number and 48 variables related to vessel and journey features. The Lloyd's data for year 2016 was prepared using the same steps above by combining several separate files purchased from Lloyd's that carried information about vessel characteristics, vessel movements, vessel types, and the places the vessels visited. Tasman Bay was renamed to Nelson, string fragments "Terminal", "Anch.", and "(NZL)", and visits to Maari Field, Maari SPM, Umuroa, Taharoa, and Westport were removed. The Lloyd's 2016 data had 9591 records with 1006 unique LMIU and 34 variables. Two Lloyd's 2000-2005 and 2016 data sets were combined and saved as a single file. Selected variables in the final file were "imo.number", "year", "gen.type", "gen.def", "place.name", and "dwt" which represented IMO number, year, generic type code, vessel type, dead weight tonnage, and the places visited by vessels.

5.2.2 MPI data set

The MPI data were provided as a single file “MPI_Vessel_data.csv” with 35235 records and 28 variables. The Lloyd’s data only contains vessel movements for vessels over 100 tonnes and therefore did not include most recreational vessels such as international yachts. The MPI data included all internationally arriving vessels irrespective of size. Therefore, it included the vessel types "Yacht”, “Yacht Catamaran”, “Yacht Trimaran", "Launch", and "Superyacht 30 Metre Plus". To be consistent with Lloyd’s data, these vessels were removed from MPI data. Data for MPI was available from 1 January 2012 until 30 April 2017, so 2017 was removed from these data. The visiting places Christchurch, Marsden Point, and Invercargill were renamed to Lyttelton, Whangarei, and Bluff, respectively. Records related to ports Otago and Port Chalmers were combined to Dunedin, and visits to Chatham Islands, Waitangi, Westport, and Whenuapai were removed. The Lloyd’s and MPI data sets prepared here were visually compared in Appendix A – section 6.3.

After assessing the similarities and discrepancies between the two datasets, it was decided to only use year 2016 data that were shared in both. To compare the datasets and for model fitting purposes, a consistent set of port names was needed that had to be contained in both datasets. While not ideal, as it potentially might cut out high risk sites, this was required to conduct further analysis using Levenshtein distance. To do so, year 2016 data for Lloyd’s and MPI data sets were matched by their common IMO numbers, which automatically excluded all non-valid IMO numbers that have been incorrectly recorded in the MPI data. It was investigated whether the vessels which could not be matched by IMO could be matched by another means. An additional six could be matched using vessel name, and there was one entry in MPI’s craft registration number column that matched with the call sign column in Lloyd’s. Ports Akaroa and Opua (previously merged with Bay of Islands) were in Lloyd’s but not in MPI data, so they were excluded. More details about comparing and matching the records of Lloyd’s and MPI datasets can be found in Appendix A – section 6.4.

5.3 Visually comparing two data sets

MPI and Lloyd’s datasets were visually compared to explore any similarity in the traffic that the New Zealand ports received each year.

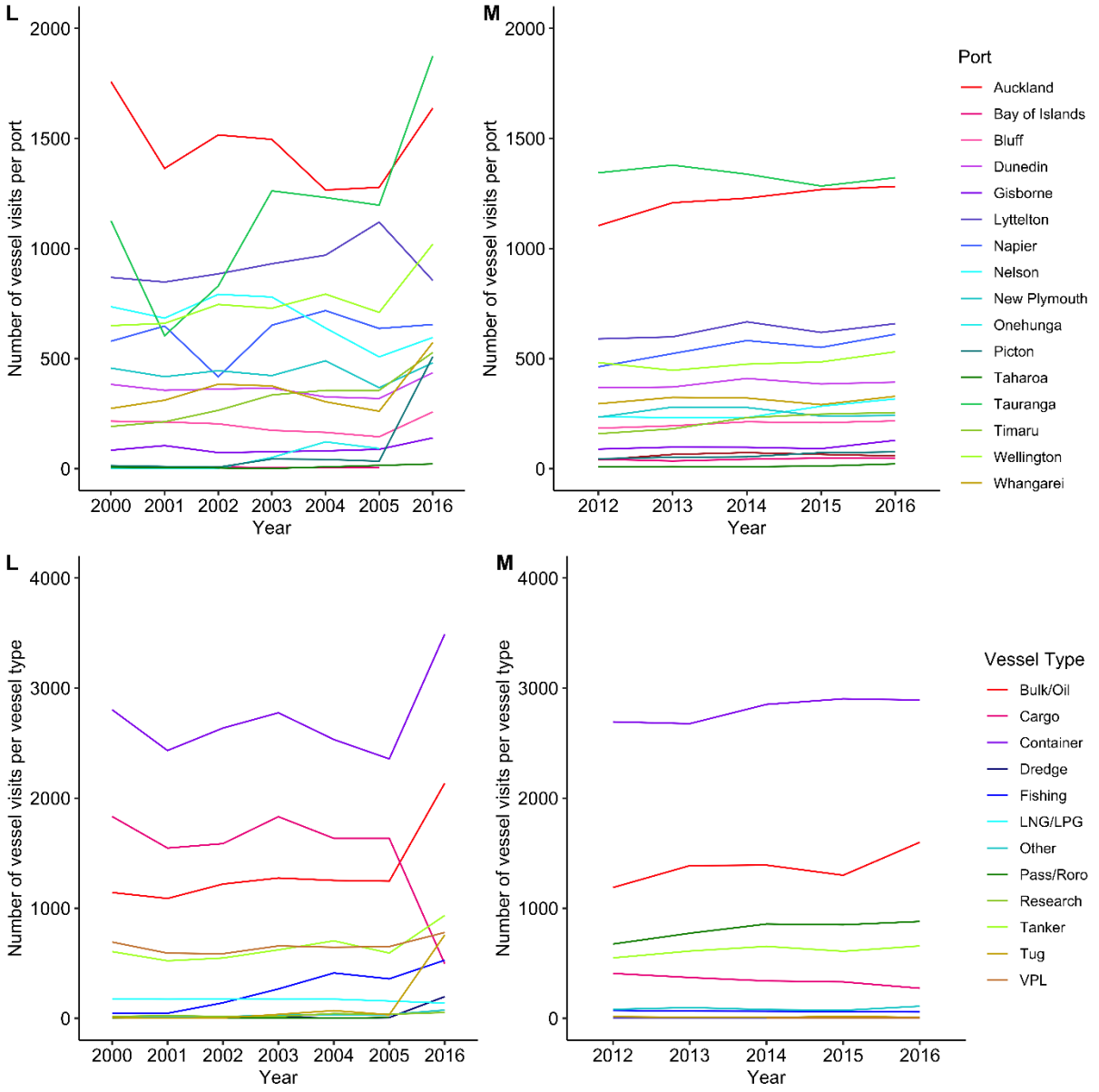

Figure 1. Total annual number of vessel visits by port (top) and vessel type (bottom). **L** represents Lloyd's 2000–2005 and 2016 data and **M** represents MPI 2012–2016 data.

Figure 1 provide direct comparisons of the number of vessel visits for each port (top) and vessel type (bottom) between Lloyd's (years 2000–2005 and 2016) and MPI (2012–2016) datasets. According to this figure, there was a substantial increase in the number of vessels arriving to New Zealand ports in the past decade, specially at Tauranga, Wellington and Whangarei. This increase was more obvious between years 2005 and 2016 in Lloyd's data mainly due to an elevation in the vessel visits at the arrival ports of Tauranga, Picton, Auckland, Whangarei, and Wellington. The number of vessel visits at Lyttelton decreased in this period (Figure 1, top left). The sudden increase in the vessel visits from 2006 to 2016 in Lloyd's data was accompanied by an increase in bulk and container vessels, although a substantial decrease in cargo vessels (Figure 1, bottom left). Tauranga has the largest number of vessel arrivals followed closely by Auckland in the MPI data (Figure 1, top right). This

corresponds with the Lloyd's 2016 data, but not the 2000–2005 data, in which Auckland had the greatest number of arrivals. The MPI database shows similar trends to the Lloyd's data, with container ships the largest arrival of all vessel types across all years, and bulk carriers second (Figure 1, bottom right).

These data are further disaggregated into port arrivals by vessel type for Lloyd's and MPI data in Figure 2 and Figure 3, respectively. The highest number of visits in Lloyd's data were by container, cargo, bulk/oil, VPL, and tanker vessels, whereas container, bulk/oil, pass/roro, tanker, and cargo had highest visits in MPI data. According to Figure 2, Auckland, followed by Tauranga, Wellington, Nelson, Lyttelton, and Napier were more visited by containers. The number of visits by containers increased in Tauranga, Picton, and Napier, but decreased in Nelson. Auckland experienced an increase in containers in 2016, following a decrease from 2000 – 2005. Auckland, Tauranga, Napier, and Lyttelton were the ports most visited by cargo carriers. Bulk/oil vessels recorded the highest number of visits after container and cargo vessels and were dominate at Tauranga, Whangarei, Auckland, Napier, and Lyttelton. As illustrated in Figure 3, similar to Lloyd's, containers had the highest number of visits and mostly arrived in large ports, namely Auckland, Tauranga, Lyttelton, Napier, Dunedin, and Wellington in MPI data. In MPI data, similar to Lloyd's, bulk/oil vessels were highest in Tauranga and Whangarei, but less frequent in Auckland than in Napier and Lyttelton. Pass/roro and tanker vessels mainly visited Auckland and New Plymouth, respectively. Cargo vessels mostly visited Tauranga, Auckland, Whangarei, Napier, and Wellington in MPI data (Figure 3).

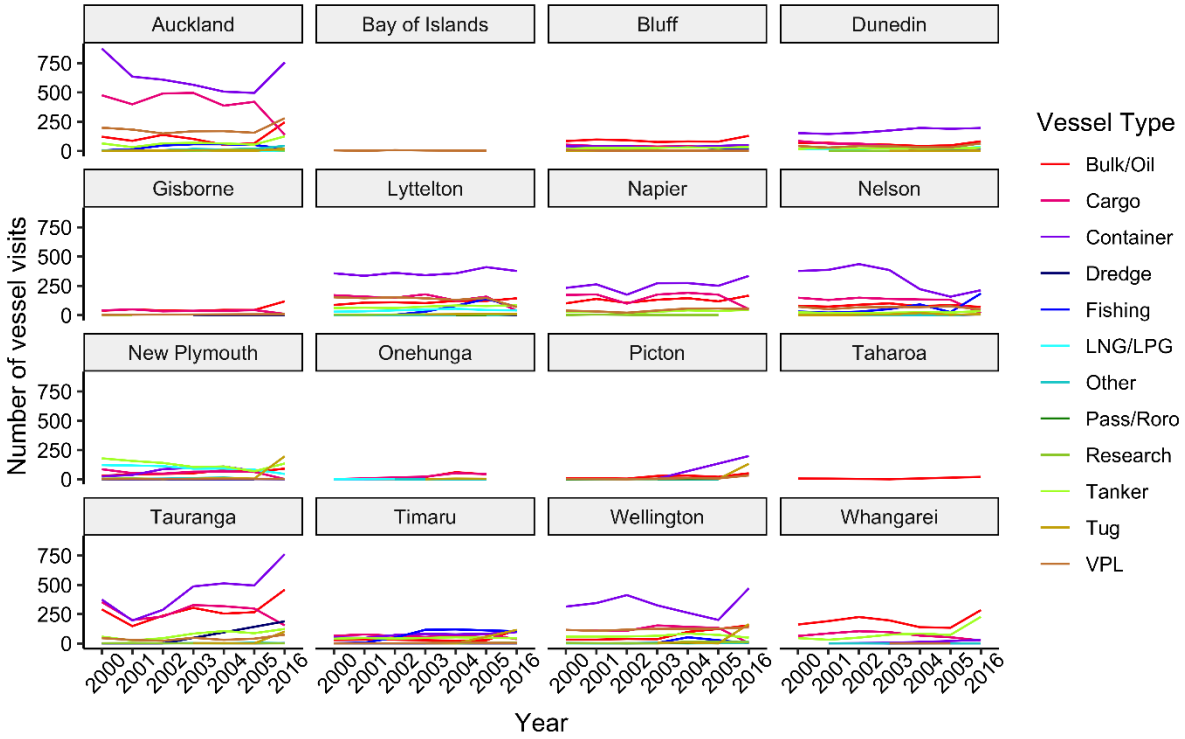

Figure 2. Comparison of the number of visits by vessel type for each port (Lloyd's 2000-2005 and 2016 data)

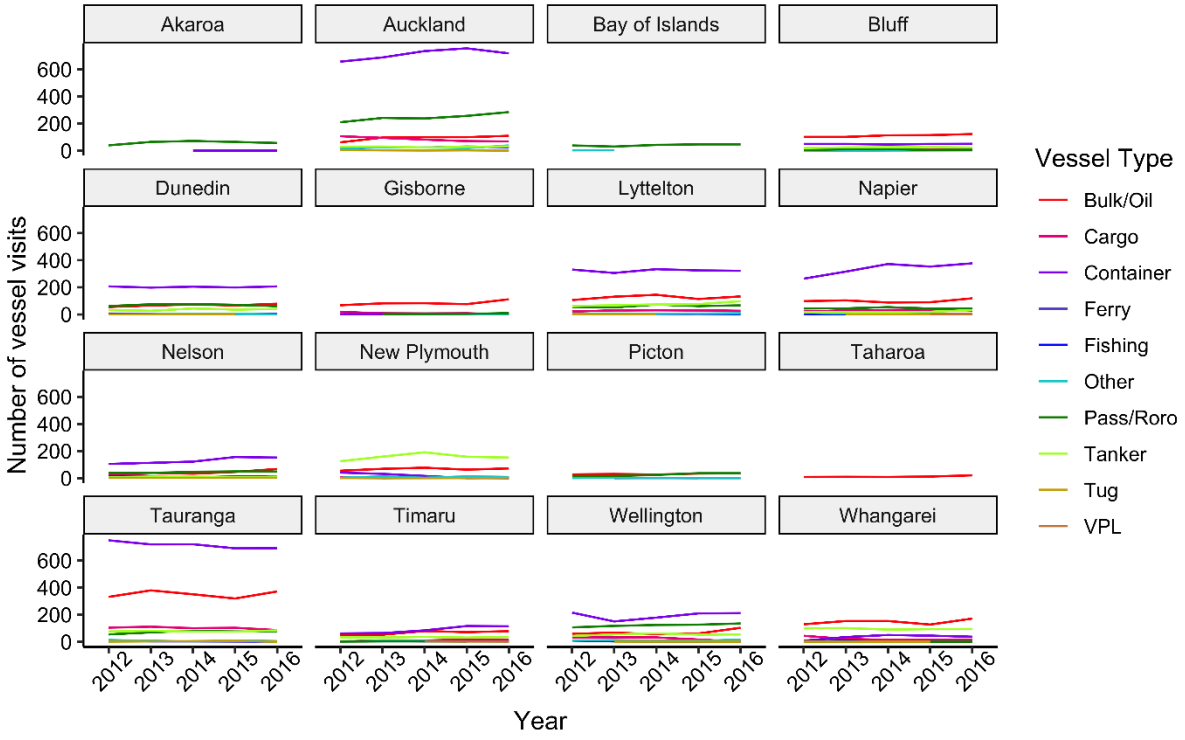

Figure 3. Comparison of the number of visits by vessel type for each port (MPI 2012–2016 data)

5.4 Comparing and matching the records

The initial analysis assessed the similarities and differences in the MPI and Lloyd's datasets. Table 2 summarises the total annual number of vessels with unique IMO number and the total annual number of port visits for both data sets. Only data for 2016 are available for

both datasets to use for comparison and the remainder of the analysis focuses on these data. After data cleaning described in Appendix A – section 6.2, and according to this table, Lloyd's 2016 data contained 1006 unique vessels making 9,591 port visits whereas there were 1,001 unique vessels in MPI data making 6,490 port visits.

Table 2. A comparison of the total number of vessels with unique IMO number and total number of port visits in Lloyd's data and MPI data for each year

Year	Total number of Vessels with unique IMO		Total number of port visits	
	Lloyd's	MPI	Lloyd's	MPI
2000	876	-	7355	-
2001	827	-	6443	-
2002	875	-	6944	-
2003	890	-	7706	-
2004	846	-	7537	-
2005	813	-	7163	-
2012	-	925	-	5685
2013	-	979	-	6000
2014	-	957	-	6253
2015	-	939	-	6152
2016	1006	1001	9591	6490

There were 922 vessels that are contained in both the MPI and Lloyd's databases for 2016 after port merging. These were matched with their individual International Maritime Organization (IMO) identifier. These vessels make 7,337 port visits in Lloyd's and 6,226 port visits in MPI data. From the 922 vessels that are contained in both databases, there are 353 vessels with a different number of port visits recorded between MPI and Lloyd's data (Figure 3). Two Cook Strait ferries with 463 and 1196 port visits caused inconsistency between two data sets. These ferries with frequent movement between Picton and Wellington were removed from the Lloyd's data. Therefore, 920 vessels remained in both data sets with 6871 port visits of common vessels for the Lloyd's data.

Table 3. The difference in port visits recorded in MPI and Lloyd's 2016 data for individual vessels

Difference in the number of port visits	Number of vessels	Vessel type
0	569	
1	202	
2	70	
3	30	
4	17	
5	4	
6	7	
7	2	
8	3	
9	2	
10	1	
11	2	
12	2	
15	2	
16	1	
17	1	
20	1	
36	1	
46	1	Dredger
66	1	NZ Flagged Product tanker
114	1	NZ Flagged Product tanker
463	1	A Cook Strait Ferry
1196	1	A Cook Strait Ferry
Grand Total	**922**	

Table 4 illustrates the number of unique and common visits for different vessel types in year 2016. Container, bulk/oil, and tankers are the vessel type with the highest number of visits in both data sets. Table 5 summarises the total number of visits by common vessels arriving in the ports merged from both data sets in 2016. Lloyd's database contains 14 New Zealand ports for the 2016 arrival data and MPI has 16 ports. To have a consistent set of port names in both data sets, ports Akaroa and Opua (previously renamed as Bay of Islands in section 6.2) were removed from the analysis.

Table 4. The total number of vessel types with unique IMO number and the total number of port visits per vessel type in Lloyd's data and MPI data for year 2016

Vessel type	Total number of Vessel types with unique IMO		Total number of port visits per vessel type	
	Lloyd's	MPI	Lloyd's	MPI
Bulk/oil	374	370	2136	1600
Cargo	83	52	496	273
Container	145	165	3490	2892
Dredge	6	-	202	-
Ferry	-	2	-	4
Fishing	43	37	528	58
LNG/LPG	8	-	139	-
Other	18	62	75	110
Pass / roro	-	138	-	880
Research	11	-	53	-
Tanker	158	164	935	658
Tug	20	7	757	10
VPL	140	4	780	5

Table 5. The total number of visits for common vessels arriving in the ports shared in both Lloyd's and MPI data in 2016

Port	Number of vessel visits for common vessels		Number of vessel visits for vessels with unique IMO	
	Lloyd's	MPI	Lloyd's	MPI
Akaroa	-	58	-	22
Bay of Islands	-	44	-	25
Auckland	1291	1204	384	360
Bluff	230	215	151	144
Dunedin	392	386	149	145
Gisborne	133	129	112	108
Lyttelton	691	626	278	258
Napier	616	581	242	225
Nelson	318	292	125	115
New Plymouth	304	238	148	140
Picton	76	72	54	52
Taharoa	21	22	3	4
Tauranga	1567	1286	496	470
Timaru	282	251	132	130
Wellington	535	509	228	230
Whangarei	415	313	243	224

5.5 GLMM model parameters and model comparison

Parameter estimates and statistics of Generalized Linear Mixed model including flag, last country, and vessel type group as fixed effect (presented in this Table 6) and IMO number as random effect (presented in Table 7) are presented here. Table 8 summarises the results of model comparison using AIC and ΔAIC.

Table 6. Parameter estimates and statistics of Generalized Linear Mixed model including flag, last country, and vessel type group as fixed effect. Coefficient estimates, standard error, confidence interval, t value, and *p*-value are presented for each predictor.

Predictors	***Estimates***	***std. Error***	***CI***	***t value***	***p-value***
(Intercept)	0.75	0.7	-0.62 – 2.13	1.07	0.283
Flag AG	-0.02	0.35	-0.70 – 0.66	-0.05	0.962
Flag AN	-0.05	0.6	-1.22 – 1.13	-0.08	0.937
Flag AU	0.37	0.5	-0.62 – 1.35	0.73	0.467
Flag BB	-0.79	0.84	-2.43 – 0.86	-0.93	0.35
Flag BM	-0.21	0.41	-1.02 – 0.60	-0.51	0.613
Flag BS	-0.01	0.32	-0.64 – 0.61	-0.05	0.964
Flag CK	-0.4	0.68	-1.73 – 0.93	-0.58	0.559
Flag CN	0.09	0.33	-0.55 – 0.73	0.27	0.787
Flag CW	0.84	0.68	-0.49 – 2.18	1.24	0.216
Flag CY	0	0.37	-0.72 – 0.73	0.01	0.993
Flag DE	1.15	0.62	-0.06 – 2.37	1.86	0.063
Flag DK	0.28	0.33	-0.38 – 0.93	0.82	0.411
Flag DM	-0.03	0.72	-1.45 – 1.39	-0.05	0.963
Flag EQ	2.63	1.06	0.55 – 4.71	2.48	**0.013**
Flag ES	0.7	0.82	-0.91 – 2.30	0.85	0.393
Flag FJ	-0.54	0.85	-2.21 – 1.13	-0.63	0.527
Flag FO	2.94	0.85	1.27 – 4.61	3.45	**0.001**
Flag GB	0.26	0.33	-0.39 – 0.91	0.8	0.425
Flag GI	0.24	0.63	-1.00 – 1.48	0.37	0.708
Flag GR	-0.08	0.43	-0.92 – 0.76	-0.19	0.852
Flag HK	0.15	0.31	-0.46 – 0.77	0.49	0.621
Flag HR	0.1	0.84	-1.54 – 1.74	0.12	0.906
Flag IM	0.09	1.14	-2.15 – 2.33	0.08	0.935
Flag IN	-0.4	0.58	-1.55 – 0.74	-0.69	0.489
Flag IT	-0.35	0.6	-1.53 – 0.83	-0.58	0.561
Flag JP	-0.3	0.47	-1.22 – 0.61	-0.65	0.518
Flag KR	0.82	0.52	-0.19 – 1.83	1.59	0.111
Flag KY	-0.21	0.37	-0.93 – 0.52	-0.56	0.576
Flag LR	0.01	0.32	-0.61 – 0.63	0.03	0.977
Flag MH	-0.04	0.31	-0.66 – 0.58	-0.13	0.9
Flag MT	-0.06	0.33	-0.71 – 0.59	-0.19	0.847
Flag NL	0.52	0.37	-0.20 – 1.24	1.41	0.157
Flag NO	-0.16	0.34	-0.83 – 0.52	-0.45	0.651
Flag NZ	2.78	0.45	1.89 – 3.66	6.12	**<0.001**
Flag PA	0.09	0.31	-0.51 – 0.70	0.3	0.762
Flag PH	-0.5	0.85	-2.16 – 1.16	-0.59	0.553
Flag PK	-0.45	0.58	-1.60 – 0.69	-0.78	0.435
Flag PT	-0.14	0.41	-0.95 – 0.66	-0.35	0.73
Flag RU	0.59	0.54	-0.48 – 1.65	1.08	0.28

Table 6 (continued)

Predictors	*Estimates*	*std. Error*	*CI*	*t value*	*p-value*
Flag SA	-0.48	0.47	-1.40 – 0.44	-1.02	0.305
Flag SE	-0.19	0.46	-1.08 – 0.71	-0.41	0.682
Flag SG	0.15	0.31	-0.46 – 0.77	0.49	0.621
Flag TO	-1.4	1.08	-3.52 – 0.71	-1.3	0.194
Flag US	0.52	0.61	-0.68 – 1.72	0.85	0.393
Flag VU	0.23	0.57	-0.89 – 1.34	0.4	0.692
Last country Argentina	-0.29	0.66	-1.58 – 1.01	-0.43	0.666
Last country Australia	-0.06	0.63	-1.30 – 1.18	-0.09	0.93
Last country Bahamas	-0.09	1.01	-2.08 – 1.89	-0.09	0.927
Last country Brazil	-0.49	0.68	-1.83 – 0.84	-0.72	0.469
Last country Brunei Darussalam	-0.69	0.71	-2.08 – 0.71	-0.96	0.335
Last country Canada	0.11	0.69	-1.24 – 1.45	0.16	0.875
Last country Chile	-0.5	0.66	-1.80 – 0.81	-0.75	0.454
Last country China	-0.22	0.63	-1.46 – 1.03	-0.34	0.733
Last country Christmas Island	0.26	0.79	-1.29 – 1.82	0.33	0.738
Last country Cook Islands	-0.03	0.71	-1.43 – 1.37	-0.04	0.97
Last country Costa Rica	0.15	1	-1.82 – 2.12	0.15	0.879
Last country Ecuador	-0.25	1.01	-2.23 – 1.72	-0.25	0.801
Last country Falkland Islands	-1.61	1.01	-3.58 – 0.36	-1.6	0.109
Last country Fiji	-0.02	0.64	-1.27 – 1.23	-0.03	0.975
Last country France	-0.77	1.03	-2.78 – 1.25	-0.75	0.456
Last country French Polynesia	0.22	0.65	-1.05 – 1.49	0.34	0.736
Last country Hong Kong	-0.03	0.64	-1.29 – 1.23	-0.05	0.964
Last country India	-0.95	0.93	-2.77 – 0.87	-1.02	0.307
Last country Indonesia	-0.03	0.64	-1.28 – 1.21	-0.05	0.958
Last country Ireland	-0.51	1.01	-2.49 – 1.48	-0.5	0.615
Last country Japan	-0.05	0.64	-1.30 – 1.20	-0.08	0.935
Last country Kiribati	-0.68	0.87	-2.39 – 1.03	-0.78	0.437
Last country Korea	-0.05	0.64	-1.30 – 1.20	-0.08	0.935
Last country Malaysia	0.05	0.65	-1.21 – 1.32	0.08	0.936
Last country Mauritius	-0.83	1.01	-2.82 – 1.15	-0.82	0.41
Last country Micronesia	-0.29	0.87	-2.00 – 1.43	-0.33	0.744
Last country Morocco	-0.85	1	-2.82 – 1.12	-0.84	0.399
Last country Nauru	-0.71	1.01	-2.69 – 1.26	-0.71	0.479
Last country New Caledonia	-0.06	0.64	-1.32 – 1.20	-0.1	0.923
Last country New Zealand	-0.07	0.73	-1.50 – 1.36	-0.1	0.924
Last country Niue	-0.36	0.88	-2.09 – 1.37	-0.4	0.686
Last country Norfolk Island	1.65	1.06	-0.42 – 3.72	1.56	0.119
Last country Northern Mariana Islands	0.25	0.74	-1.20 – 1.71	0.34	0.733
Last country Oman	-0.81	0.74	-2.26 – 0.65	-1.09	0.277
Last country Panama	0.03	0.64	-1.21 – 1.28	0.05	0.959

Table 6 (continued)

Predictors	*Estimates*	*std. Error*	*CI*	*t value*	*p-value*
Last country Papua New Guinea	-0.28	0.72	-1.69 – 1.13	-0.39	0.694
Last country Peru	-0.19	0.93	-2.01 – 1.63	-0.2	0.838
Last country Philippines	-0.06	0.64	-1.32 – 1.20	-0.09	0.927
Last country Portugal	-0.57	1.01	-2.54 – 1.41	-0.56	0.573
Last country Qatar	-0.49	0.9	-2.24 – 1.27	-0.55	0.586
Last country Russia	-0.35	0.84	-2.00 – 1.30	-0.42	0.674
Last country Samoa	-0.61	0.74	-2.05 – 0.83	-0.83	0.409
Last country Samoa, American	-0.29	0.67	-1.61 – 1.03	-0.44	0.663
Last country Saudi Arabia	-0.4	0.69	-1.75 – 0.95	-0.58	0.561
Last country Singapore	0.15	0.64	-1.10 – 1.40	0.23	0.817
Last country Solomon Islands	0.1	0.69	-1.25 – 1.45	0.14	0.885
Last country South Africa	0.31	0.79	-1.23 – 1.85	0.39	0.694
Last country Spain	-0.23	0.75	-1.71 – 1.25	-0.31	0.758
Last country Sri Lanka	0.77	1.01	-1.20 – 2.74	0.76	0.445
Last country Taiwan	-0.21	0.75	-1.67 – 1.26	-0.27	0.784
Last country Thailand	0.35	0.82	-1.25 – 1.95	0.43	0.669
Last country Tokelau	2.15	1.12	-0.05 – 4.35	1.91	0.056
Last country Tonga	-0.02	0.65	-1.30 – 1.25	-0.04	0.971
Last country United Arab Emirates	-0.46	0.66	-1.75 – 0.83	-0.7	0.482
Last country United States of America	-0.03	0.64	-1.28 – 1.22	-0.05	0.959
Last country Unknown	-0.4	0.75	-1.86 – 1.07	-0.53	0.597
Last country Vanuatu	0.09	0.69	-1.27 – 1.45	0.13	0.894
Last country Vietnam	-0.39	0.67	-1.71 – 0.93	-0.58	0.562
Craft Type Grouped Container	-0.36	0.08	-0.52 – -0.20	-4.38	**<0.001**
Craft Type Grouped Cruise Liner	0.02	0.16	-0.29 – 0.32	0.1	0.92
Craft Type Grouped Fishing	-0.41	0.33	-1.06 – 0.25	-1.22	0.223
Craft Type Grouped General Cargo	-0.51	0.12	-0.75 – -0.27	-4.11	**<0.001**
Craft Type Grouped Naval	-2.08	0.53	-3.12 – -1.04	-3.92	**<0.001**
Craft Type Grouped Other	-0.19	0.14	-0.47 – 0.08	-1.4	0.161
Craft Type Grouped Roll On/Roll Off	-0.65	0.1	-0.84 – -0.45	-6.56	**<0.001**
Craft Type Grouped Tanker	-0.02	0.08	-0.18 – 0.13	-0.28	0.776

Table 7. The GLMM estimates for random effect IMO with 2313 observations and 908 IMO numbers, $\sigma^2 = 0.31$, τ_{00} (random-intercept-variance, or between-subject-variance) = 0.3, ICC (Intraclass Correlation Coefficient) = 0.49, Marginal $^{R2} = 0.16$, and Conditional $R^2 = 0.57$

Groups	**Name**	**Variance**	**Std.Dev.**
IMO	(Intercept)	0.3	0.54
Residual		0.31	0.56

Table 8. Model comparison using AIC and ΔAIC. Null and selected models are highlighted in red.

Model			Covariates			
	AIC	ΔAIC	Flag	Last country	Vessel type	First port
1	9398.91	0.00	X	X	X	
2	9402.02	3.11	X	X	X	X
3	9428.46	29.55	X		X	
4	9519.10	120.18	X			
5	9714.71	315.80		X		
6	9733.06	334.15			X	
7	9769.55	370.64				X
Null	9750.22	351.31				

References

Arthur, T., Arrowsmith, A., Parsons, S., Summerson, R. 2015a. Monitoring for Marine Pests: A review of the design and use of Australia's National Monitoring Strategy and identification of possible improvements. *Department of Agriculture and Water Resources, Canberra*.

Arthur, T., Summerson, R., Mazur, K. 2015b. A comparison of the costs and effectiveness of prevention, eradication, containment and asset protection of invasive marine species incursions. in: *ABARES Report,*, (Ed.) A.B.o.A.a.R.E.a. Sciences. Canberra.

Bolam, F.C., Grainger, M.J., Mengersen, K.L., Stewart, G.B., Sutherland, W.J., Runge, M.C., McGowan, P.J. 2019. Using the value of information to improve conservation decision making. *Biological Reviews*, **94**(2), 629-647.

Hatami, R., Lane, S., Robinson, A., Inglis, G., Todd-Jones, C., Seaward, K. 2021. Improving New Zealand's marine biosecurity surveillance programme: A statistical review of biosecurity vectors. Biosecurity New Zealand Technical Paper, (Ed.) M.f.P. Industries. New Zealand.

Heath, A., Manolopoulou, I., Baio, G. 2016. Estimating the expected value of partial perfect information in health economic evaluations using integrated nested Laplace approximation. *Statistics in medicine*, **35**(23), 4264-4280.

Inglis, G. 2018. Optimising New Zealand's marine biosecurity surveillance programme: Project Management Plan, (Ed.) M.f.P. Industries. New Zealand.

Lazreg, M.B., Goodwin, M., Granmo, O.-C. 2020. Combining a context aware neural network with a denoising autoencoder for measuring string similarities. *Computer Speech & Language*, **60**, 101028.

Levenshtein, V.I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*. Soviet Union. pp. 707-710.

Lockwood, J.L., Cassey, P., Blackburn, T. 2005. The role of propagule pressure in explaining species invasions. *Trends in Ecology and Evolution*, **20**(5), 223-228.

Lonsdale, W.M. 1999. Global patterns of plant invasions and the concept of invasibility. *Ecology*, **80**(5), 1522-1537.

Marine Environmental Protection Committee. 2018. MEPC 72/17 - Report of the Marine Environment Protection Committee on its Seventy-Second Session. International Maritime Organization.

McDonald, J.I., Wellington, C.M., Coupland, G.T., Pedersen, D., Kitchen, B., Bridgwood, S.D., Hewitt, M., Duggan, R., Abdo, D.A. 2020. A united front against marine

invaders: Developing a cost-effective marine biosecurity surveillance partnership between government and industry. *Journal of Applied Ecology*, **57**(1), 77-84.

Morisio, M., Baccaglini, E., Brevi, D., TesfayeAlemu, G. 2018. Analysis And processing Of Information Transmitted By Vessels.

Ooms, J. 2017. pdftools: Text extraction, rendering and converting of PDF documents. R package version 1 ed, pp. 2017.

Shannon, C., Stebbing, P.D., Dunn, A.M., Quinn, C.H. 2020. Getting on board with biosecurity: Evaluating the effectiveness of marine invasive alien species biosecurity policy for England and Wales. *Marine Policy*, **122**, 104275.

van der Loo, P.J.M. 2014. The stringdist package for approximate string matching. *R Journal*, **6**(1), 111-122.

Wickham, H. 2019. Stringr: Simple, consistent wrappers for common string operations. R package version 1.4.0 ed.