# Supplementary Materials

## Part 1 - Data preparation

## Data preparation for the Historic Ballast Water Data

The process for cleaning and preparing the historical data for analysis as written in the R script is described in this section.

**Joining the data:** The first step of data cleaning was to combine two available data sets, one containing voyage and intended ballast discharge quantity information, and the other containing survey information, which provided the intention to discharge. The latter variable was related to the answer to the pre-arrival question of how vessels that are carrying ballast water would comply with New Zealand’s ballast water controls. The answers were: 1- No, not discharging any ballast water in New Zealand waters. 2- Yes, exchanging the ballast water mid-ocean in all tanks that are to be discharged in New Zealand waters. 3- Yes, discharging only fresh water. The created data set comprised 17755 observations, with 10798 rows containing missing values for ballast water volume. For model building, some vessel types that were uncommon in the data sets were merged with a similar class (see Table 1).

**Checking the reliability of variable intent:** For each voyage, it was checked whether there was evidence that the agreement between the stated intention to discharge and discharge volume reported varied by arrival port and vessel type. Table 3 and Table 4 suggest that the level of agreement depended upon both variables. For example, of the 8054 journeys that arrived at Auckland, only 2297 instances of discharge happened when the vessels had declared the intention of discharge, whereas 222 out of 364 instances of discharge arriving at Nelson had answered "YES" to the intention of discharge. Similarly, the answers to the intention of discharge were not consistent across different types of vessels. For example, almost half of the tanker and cargo discharged with the intention of discharge, but most (4816 out of 6607) discharges from? unitised/container carriers were not in agreement with the intention of discharge. Further exploration indicated that intention to discharge was missing for most years; consequently, this variable was removed from the analysis (Supplementary Materials – part 2).

**Selecting the ports of interest:** Sixteen ports of interest were selected: namely Auckland, Bluff, Gisborne, Lyttelton/Christchurch, Whangarei, Tauranga, Milford Sound, Napier, Nelson, New Plymouth, Picton, Dunedin, Taharoa, Timaru, Wellington, and Westport. This selection was done because the ports outside of this set were not in New Zealand, so the discharge amount was not of interest. In this dataset, 10704 visits had missing values for ballast water volume. These missing values were replaced by zeros, proscribing that non entry meant that discharge did not happen.

**Constructing the historical dataset:** The final historical dataset for ballast water analysis was built by combining variables arrival date, vessel type, deadweight tonnage (DWT), discharge port, discharge volume, arrival port, and intent of discharge. A description of these variables is provided in Table 1.

## Data preparation for contemporary data for ballast water

To build the data used to predict discharge by recent (2015-2017) vessel voyages, four data sets were merged together. The first file contained 5986 vessel arrival records matched by MPI. The other files contained additional records for ‘Coastwise’ arrivals of vessels at New Zealand ports (n = 9002) and arrivals at places of first arrival (n = 461) that were matched from the supplied data. Selected variables were arrival date, arrival port, vessel type, and DWT. This data set was also used as the contemporary data for the biofouling exposure analysis. The contemporary data set for ballast analysis had 15172 observations, with 63, 64, and 1085 missing values for the variables DWT, vessel age, and vessel speed, respectively.

## Part 2 - Agreement between the stated intention to discharge and discharge volume

We explored whether the number of discharge instances at each port agreed with the number of times that the answer to the intention of the discharge was "Yes". Figure 1 illustrates the number of discharge instances in the discharge port by arrival port based on the intentions of ballast discharge. Part 1 of the graph shows the number of instances of ballast water discharge when the answer to the intention of the discharge was "NO", implying not discharging any ballast water in New Zealand waters. Parts 2 and 3 of the figure show the instances of ballast water discharge when the answer was "YES". The size of the circles in the graph represents the counts of discharge instances. That is, the larger the circle, the higher the number of discharges of ballast water. The circles are presented in yellow colour when the discharge port is the same as the arrival port, and they are shown in pink when the discharge port is unknown.

To prepare data for this graph, a few changes was made to the data preparation. Omitting the missing values from the ballast data set with 17755 observations reduced the number of observations to 6114 records, almost all of which had non-zero ballast discharge. Therefore, it was not possible to identify which voyages did not involve ballast discharge. In addition, by removing these missing data, it would be difficult predicting the discharge ports for some of the arrival ports in the contemporary data, because there were no discharge ports for some of the arrival ports in the historical data. This problem was solved by not removing the NA values from the historical data set, and a few other modifications. Instead, the ports that are not among the discharge ports of interest were named "FOREIGN" (as illustrated in Figure 1). For example, from the vessels that arrived at Auckland and claimed not discharging ballast water, 6015 vessels discharged at unknown ports, and 139 vessels discharged at FOREIGN ports, i.e., one of the ports that were not included in the list of our ports of interest. From the vessels that arrived at Tauranga, 1908 discharged at unknown ports, and 41 vessels discharged at FOREIGN ports. One important feature of this graph is the higher number of discharges where there was no intention of discharge (part 1). From this graph, it seemed that there was not an agreement between discharge and the intention of discharge. For example, 738 instances of discharge occurred when vessels that arrived in Auckland did not intend to discharge, but they specified a proposed discharge volume, and it was not clear where the discharge happened.

After finding evidence that the first arrival port and vessel type affected this agreement (Table 3 and Table 4), generalized linear mixed models were fitted to test this effect, and any potential interaction effect. To do so, a new data frame was made with a variable "correct answers" which shows agreement between discharge and intention of discharge. The outcome of the binomial GLM showed that the proportion of correct answers to total answers was strongly affected by vessel type and arrival port as follows:

In this equation, is an indicator for the correct answers, is the probability of correct answers, and is the logit link function. Parameters and are the arrival port and vessel type effects for the probability of correct answers.

Table 1. The description of the variables used in the historical and contemporary data set for ballast water analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **variable name** | **Description in historical data** | **Description in contemporary data** |
| 1 | Arrival date | The date the vessel arrived and covered the years 1998-2008. | The date the vessel arrived and covered the years 2015-2017. |
| 2 | Vessel type | There were 14 levels of B, F, G, L, M, MPR, N, P, R, T, U, X, Y, and O representing bulk carrier, fishing, general cargo, Liquefied Petroleum Gas Carrier (LNG/LPG), vehicle/livestock carrier, passenger/roro, naval vessel, passenger, research, tanker, unitised/container carrier, tug, drill ship, and others, respectively. Some of the vessel subtypes were combined into the other types, e.g., vessel type C was combined to type B, and type PR to type MPR. That is, all types of bulk (e.g., bulk, bulk/oil, ore/oil) were combined together, and all types of passenger (e.g., passenger, passenger ro/ro) were combined together. These categories are from the Lloyds VESSELS table derived from ([Lloyds Register 2018](#_ENREF_30)).  For brevity, unitised/container carrier and general cargo were called “container” and “cargo” in this paper, and other vessel types were mentioned without the word “carrier”. | Vessel type "D" was changed to "O", and vessel type "P" was renamed to "MPR". For all vessel type "tug”, discharge ports were in missing data for all the vessels, for this reason and as Tugs are unlikely to carry ballast water, vessel type "tug" was removed from the analysis. For the sake of brevity, general cargo was named as cargo and bulk, vehicle/livestock, container, LNG/LPG were mentioned without “carrier” in the text. |
| 3 | DWT | Dead Weight Tonnage of the vessel. | - |
| 4 | Discharged port | The port/ports in which the vessel discharged the ballast water. | No value for this variable in contemporary data set. |
| 5 | Discharged volume | The sum of the ballast water volume discharged at each discharge port. | No value for this variable in contemporary data set. |
| 6 | Arrival port | The first port that the vessel visited. Some of the ports were renamed; for example, ports Marsden Point, Tiwai Point, Mount Maunganui were renamed to Whangarei, Bluff, and Tauranga, respectively. | Ports Marsden Point, Tiwai Point, and "Mount Maunganui" were renamed to Ports Whangarei, Bluff, and Tauranga, respectively. Records related to ports Dunedin and Chalmers were combined to Port Chalmers/Dunedin, and observations related to ports Lyttelton and Christchurch were combined to port Lyttelton/Christchurch. |
| 7 | Intent | The answer to the question of the intention of discharge. The answer was "no" if there was no intention of discharge, and "yes", otherwise. | No value for this variable in contemporary data set. |
| 8 | VoyageID | - | - |
| 9 | VesselID | - | - |

Table 2. Description of the variables used for biofouling analysis.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ID** | **Predictor variable** | **Data range in historical data** | **Data range in contemporary data** | **No. NA-Historical** | **No. NA-Contemporary** | **Description** |
| 1 | Antifouling age | 3 - 1946 | 5.83 - 43090 | 15 | 0 | Time (days) since last dry-docked or new antifouling applied. |
| 2 | Vessel type | - | - | 0 | 0 | There are 17 types of vessels. Types were Barge, Bulk Carrier, Bulk Craft, Container, Cruise Liner, Ferry, Fishing, Cargo, Livestock Carrier, Naval, Oil Supply, Roll, Specialised Craft, Superyacht, Tanker, Tug, and Yacht. |
| 3 | Vessel age | 0 - 49 | 1 - 113 | 0 | 64 | Vessel age (year) in contemporary data was calculated as the difference between the date of arrival in port and the date the vessel was built. |
| 4 | Vessel speed | 8 - 28.5 | 9.5 - 29 | 9 | 1085 | The design speed of the vessel. |
| 5 | Duration of maximum lay-up | 0 - 82 | 9 - 589 | 104 | 0 | Maximum duration of lay-up (days) since last dry deck or antifouling. MMDays from historical data was transformed into a categorical variable with 4 bins: <10, 10 x 20, 20 x 30, 30 and used these to build MMDays in contemporary data. |
| 6 | Month | - | - | 0 | 0 | Season or month of arrival |
| 7 | Mass | 0 - 1251.44 |  |  |  |  |

Table 3. Agreement between discharge and intention of discharge varied by arrival port.

|  |  |  |
| --- | --- | --- |
| **First arrival port** | **Agreement** | |
| **True** | **False** |
| Auckland | 2297 | 5757 |
| Bluff | 278 | 276 |
| Chatham Islands | 1 | 1 |
| Gisborne | 86 | 58 |
| Lyttelton/Christchurch | 298 | 325 |
| Milford Sound | 17 | 30 |
| N/R | 9 | 13 |
| Napier | 205 | 191 |
| Nelson | 222 | 142 |
| New Plymouth | 749 | 389 |
| Picton | 23 | 9 |
| Port Chalmers/Dunedin | 74 | 243 |
| Taharoa | 38 | 7 |
| Tauranga | 808 | 1856 |
| Timaru | 56 | 65 |
| Wellington | 113 | 196 |
| Westport | 2 | 6 |
| Whangarei | 423 | 392 |
| Unknown | 1 | 0 |

Table 4. Agreement between discharge and intention of discharge varied by vessel type.

|  |  |  |
| --- | --- | --- |
| **Vessel type** | **Agreement** | |
| **True** | **False** |
| Bulk carrier (B) | 1575 | 1023 |
| Cargo (C) | 2 | 3 |
| Fishing (F) | 3 | 31 |
| General cargo (G) | 1041 | 2014 |
| Liquefied Petroleum Gas Carrier (L) | 129 | 145 |
| vehicle carrier/livestock (M) | 369 | 1213 |
| Naval vessel (N) | 2 | 11 |
| Passenger (P) | 0 | 4 |
| Research (R) | 14 | 32 |
| Tanker (T) | 758 | 606 |
| Unitised / container carrier (U) | 1791 | 4816 |
| Tug (X) | 4 | 18 |
| drill ship (Y) | 0 | 2 |
| Others (O) | 12 | 38 |



Figure 1. Number of discharge instances in ports based on the intention to discharge ballast. Data are presented by the first port of arrival. The green colour represents the number of ballast water discharge instances at any port except for the arrival port. The yellow colour shows the number of ballast water discharge instances for vessels that used their arrival port as discharge port/ports. The pink circles represent the number of ballast water discharge instances related to the ports that the discharge port for which is not defined (NA). FOREIGN refers to the ports that are not among the discharge ports of interest.

## Part 3 - Estimating the total wetted surface area (TWSA) and the area of niches for each arriving vessel

The total mass of biofouling on a vessel was estimated as the product of its estimated density and the total area fouled. As most biofouling occurs in niche areas of vessels, we estimated the total area of niches on each vessel using relationships developed by [Moser et al. (2016)](#_ENREF_41) and [Moser et al. (2017)](#_ENREF_40).

[Moser et al. (2016)](#_ENREF_41) showed that the TWSA of a vessel can be estimated accurately by:

where DWT is the Dead Weight Tonnage of the vessel and the coefficient, , and exponent, , describe the shape of the regression for each class of merchant vessel. Values for and are given in Table 2 of [Moser et al. (2016)](#_ENREF_41) for each vessel class. Data on the DWT of each vessel and its class type were contained in the dataset acquired from Lloyds Maritime Intelligence Unit for this study and were cross-checked against data supplied to the project by MPI. In the latter, only data on the Gross Tonnage (GT) of the vessel were available. However, DWT can be estimated from GT using linear regression models specified in Table 3 of [Moser et al. (2016)](#_ENREF_41).

The area of niches on a vessel can be estimated simply as a percentage of the TWSA. Estimates of the percentage of TWSA comprised by niches for different classes of vessel are provided in Table 4 of [Moser et al. (2017)](#_ENREF_40)

## Part 4 – More details about KNN analysis algorithm

Whenever a vessel did not discharge in a port, it was assumed that the lack of discharge was recorded as a 0 discharge. The data was rearranged to take this assumption into account as provided in Figure 2.

A picture containing chart

Description automatically generated

Figure 2. Example of transforming voyage event data to wide format.

Figure 3 shows a schematic of cross-validation. The top row shows the original data; circles are outcomes, squares are predictors, and the colours represent different values. This particular type of cross-validation is known as K-fold cross-validation: the original data is split into K groups (the example in Figure 3 shows three groups); K-1 groups are used for training (fitting) models, with the remaining group being used for testing/comparing predictions. As already discussed, each group for the ballast water application is a single year of voyages.

Consider the cross-validation group #A from Figure 3; the models are built using the magenta, orange, green, and pink predictors. In this example, we fit *l* different models, indexed by the hyperparameters *θ1, …, θl*; each model results in predictions , …, . The hyperparameters in our example are the categorical variables (i.e., whether to group by vessel type or port of arrival, or both), the numeric variables (using dead weight tonnage and/or date of arrival in the nearest neighbour matrix) and the number of nearest neighbours. We repeat this process (using the same hyperparameter sets) within each of the *K* folds, resulting in predictions indexed by the folds (A, B, C in this example). Note that each outcome y is only used once for predictions.



Figure 3. An example of cross-validation. The top row shows the original data; circles are outcomes, squares are predictors. Each cross-validation fold (CV group) is used to fit the model, with the predictions shown on the right-hand side.

The hyperparameters are chosen so that the cross-validation error in the predictions is minimised. To do this, we define a loss function which we then apply to each of the held-out testing sets (years) and average the results. For example, for the model with hyperparameters *θl*, and folds as shown in Figure 3, this is . We then choose the set of hyperparameters that gives the minimum L.

## Part 5 - Exploratory analysis

The results of exploratory analysis revealing the variations in the number of visits, discharge volume, discharge counts, and DWT for the vessels arriving at New Zealand ports and discharging ballast water in discharge ports during years 1999 – 2017 are given in this section. Generally, the annual total number of visits increased in contemporary data and reached the highest record in 2016. Auckland was the busiest port until it was surpassed by Tauranga in 2016. Other busiest ports were Lyttelton, Wellington, Napier, and Nelson. The increase in arrivals during 2015 – 2017 was accompanied by an increase in the number of visits by vessel types of container, cargo, bulk/oil, tanker, and vehicle/livestock. These vessels had the highest number of visits in 1999 – 2017. Vessel types container and bulk/oil, followed by cargo and tanker had the highest vessel discharge counts and DWT. Auckland followed by Tauranga, Lyttelton, Napier, and Wellington had the largest total DWT during 1999 – 2007; whereas in the contemporary data, Tauranga, Auckland, Lyttelton, Napier, and Whangarei had the highest DWT level, respectively. Although Auckland was the busiest arrival port in the historical data, the highest ballast water volume was discharged in New Plymouth, Tauranga, and Lyttleton ports. Tauranga as a discharge port had the highest values of vessel discharge counts and DWT for most years, as well as the highest number of visits and DWT as an arrival port after Auckland. Vessels discharged the highest volume of ballast water at their arrival port. In this case, New Plymouth, Lyttelton, Napier, and Tauranga, followed by Taharoa and Whangarei received the highest discharge.

Figure 4 – Figure 8, Figure 11 and Figure 12 provide direct comparisons of vessel visits and DWT between the historical (years 2000–2005) and contemporary (2016) Lloyd’s datasets. Lloyd’s dataset was used because the historic ballast dataset did not have records for all arrivals over 1999-2007. These summaries do not affect the predictive modelling and its interpretation, and this is for describing the changes in shipping to NZ over the past decade. Figure 9, Figure 10, Figure 13 – Figure 15 illustrate the variations in the discharge volume, discharge counts, and DWT for the vessels arriving at New Zealand ports and discharging ballast water during years 1999–2007 using MPI datasets. Data related to the years 1998 and 2008 only included partial counts for the year and were removed from the analysis. Where Lloyd’s data are available, these are used.

Figure 4 shows the total annual number of visits across all vessels and ports during years 2000–2005 and 2016, for example, one vessel visiting three ports would be counted as three visits. From this figure, the total annual number of visits was highest in 2016.



Figure 4. Total annual number of visits across all vessels and ports from 2000–2005 and 2016 (Lloyd’s data).

According to Figure 5, this increase was due to an elevation in the traffic at the arrival ports of Auckland, Tauranga, Wellington, Napier, and Nelson. Auckland had the largest number of visits during 2000–2005, followed by Tauranga (except for 2001 and 2002), however, this reversed by 2016. Except for Lyttelton that experienced a decrease in the traffic, almost all other ports received a higher number of visits in 2016.



Figure 5. Total annual number of visits for each port for historical and contemporary data during years 2000–2005 and 2016 (Lloyd’s data). Vertical line highlights gap in data from 2006–2015.

Figure 6 shows the number of arriving vessels by type and year across all ports. The vessel types with high number of visits were container, cargo, bulk/oil, tanker, and vehicle/livestock, with container having the highest number of visits across all years. As the number of vessels visiting New Zealand ports increased from 2001 to 2003 (Figure 4), so did the total number of vessel types container, cargo, bulk/oil, tanker, and vehicle/livestock carrier. The increase in arrivals in 2016 was related to containers, tankers, and bulk/oil carriers, with a reduction in cargo vessel types (Figure 6).

Figure 7 shows the total DWT of vessel types visiting New Zealand ports during years 2000 – 2005 and 2016. Container, bulk/oil, cargo, and tanker vessel types always had the highest DWT in the historical data. All vessel types except for cargo had a higher DWT in 2016 that corresponded to an increase in port visits for that year (Figure 4).



Figure 6. Annual number of visits of each vessel type across all ports for historical and contemporary data during years 2000–2005 and 2016 (Lloyd’s data). Vertical line highlights gap in data from 2006–2015.



Figure 7. Total annual DWT for different vessel types for historical and contemporary data during years 2000–2005 and 2016 (Lloyd’s data). Vertical line highlights gap in data from 2006–2015.

Vessels visiting Auckland had the largest total DWT during 2000–2002, with Tauranga overtaking during 2003 – 2005 (Figure 8). Lyttelton, Napier, and Wellington were also among the ports with the highest DWT in the historical data, whereas in the contemporary data (2016), Tauranga, Auckland, Lyttelton, Napier, and Whangarei had the highest DWT level, respectively. There was a substantial increase in the DWT in the 2016 data. Almost all the arrival ports received the highest number of visits and DWT in 2016.

****

Figure 8. Total annual DWT by port for historical and contemporary data during years 2000–2005 and 2016 (Lloyd’s data). Vertical line highlights gap in data from 2006–2015.

The vessel types; container and bulk/oil, followed by cargo and tanker had the highest vessel ballast water discharge counts and DWT in historical data; the order was the same for DWT in 2016, except for vehicle/livestock vessel that took over cargo vessels (Figure 7 and Figure 9). Ballast water discharge data are not available for the contemporary records (2016).

Bulk/oil vessel types had the highest ballast water discharge counts, with a maximum in 2002, and was surpassed by container vessels from 2003–2007 (Figure 9). Vessel type container, followed by bulk/oil also had the highest DWT during 2000–2005, and 2016 (Figure 7). Vessel discharge counts of tanker vessels reached the maximum levels in 2001 and experienced a gradual decrease afterwards. Vessel discharge counts for cargo generally increased from 1999–2005, before declining slightly in 2006 (Figure 9). While the DWT did not show a noticeable change for tanker and cargo vessels during 2000–2005, the former increased and the later decreased sharply in 2016 (Figure 7).



Figure 9. Total number of ballast water discharge counts per vessel type at each discharge port during 1999–2007 (MPI data).

Figure 10 shows the total discharge volume of different vessel types visiting discharge ports. Bulk/oil vessels discharged the highest volume of ballast water, with an increase in volume from 1999 to 2002, followed by a decrease in volume from 2002 to 2007. Tankers also had a high volume of discharged ballast water, followed by container and cargo vessels (Figure 10). Tanker vessels discharged the maximum ballast water in 2001 and 2002. The declared volumes discharged generally declined during the period 2003–2007. Overall, the discharge volume of container and cargo vessels had an upward trend during 1999–2007.



Figure 10. Annual ballast water discharge volume (tonnes) of vessel types visiting discharge ports during 1999–2007 (MPI data).

The total annual number of vessel visits and total annual DWT of each vessel type arriving at each port during years 2000–2005 and 2016 are shown in Figure 11 and Figure 12.

Auckland had a much higher number of container vessels visits compared with other ports during 2000–2005, with Tauranga the second most visited over this period (Figure 11). Container vessel visits in 2016 was highest in Tauranga followed by Auckland, Wellington, Lyttelton, Napier, and Nelson. Tauranga was also the most visited port by bulk/oil vessels from 2000–2005. Other sites most visited by bulk/oil vessels in historical data (2000–2005) were Whangarei, Napier, Lyttelton, Auckland, Bluff, and Nelson.



Figure 11. The annual number of vessels of each type arriving at each port, for historical and contemporary data during years 2000–2005 and 2016 (Lloyd’s data). Vertical line highlights gap in data from 2006–2015.

The number of bulk/oil vessels increased dramatically in 2016 compared with the earlier years, and this increase was most notable in Tauranga. Bulk/oil vessels in 2016 mostly visited Tauranga, Whangarei, Auckland, Napier, Wellington, Lyttelton, and Bluff. Cargo vessel visits during 2000–2005 were highest in Auckland, Tauranga, and Napier. In 2016, the number of cargo vessel visits decreased in all ports, especially in Auckland and Tauranga, which ended with similar visit numbers in 2016. Despite this decrease, Auckland and Tauranga still remained preferred destinations for cargo vessels, with a substantial number of these vessels also visiting Lyttelton and Napier (Figure 11).

The total annual DWT of each vessel type arriving at each port during 2000–2005 and 2016 is shown in Figure 12. Container, bulk/oil, and tanker vessels had the highest annual DWT, and there was a dramatic increase in the annual DWT of these vessels in 2016 in almost all the ports. Auckland followed by Tauranga had the highest annual DWT of container and cargo vessels during 2000–2005, and Tauranga surpassed Auckland for both vessel types in 2016. Container vessels mostly visited Tauranga, Auckland, Lyttelton, and Napier during 2016 (Figure 11). The highest annual DWT of bulk/oil vessels in historical data were observed in Tauranga, Whangarei and Lyttelton. As in the historical data, bulk/oil vessels had the next highest total DWT and mostly arrived at Tauranga in 2016 (Figure 12).



Figure 12. The annual total DWT of vessels of each type arriving at each port, for historical and contemporary data during years 2000–2005 and 2016 (Lloyd’s data), respectively. Vertical line highlights gap in data from 2006–2015.

Figure 13 – Figure 15 show the total number of ballast water discharge counts, total DWT, and total ballast water discharge volume per year for each discharge port during 1999 – 2007, respectively.

From Figure 13, it can be seen that New Plymouth, as a discharge port, was the most visited port for most of the years during 1999–2002. After this period, the number of vessels discharging at this port reduced dramatically. This was followed by an increase in the number of vessels discharging in Auckland and Tauranga in 2003–2007. Reduction in the number of vessels discharging at New Plymouth in 2003 coincided with the reduction in vessel visits across all vessel types (Figure 11).

Tauranga, Auckland, and Plymouth had the highest DWT as discharge ports during 2003–2007 (Figure 14). This figure shows a similar negative trend for DWT for New Plymouth after 2003 as seen for the number of discharge counts from Figure 13. DWT were also high at other discharge ports such as Napier and Lyttelton. At Lyttelton, DWT increased to its highest in 2006, similar to that observed in Auckland the same year.



Figure 13. Total number of vessel ballast water discharge counts at each discharge port during 1999–2007 (MPI data). FOREIGN refers to the ports that were not among the discharge ports of interest.

Although Auckland was the most visited arrival port (Figure 5), other ports such as New Plymouth, Tauranga, and Lyttelton received the highest discharge volumes per year (Figure 15). The highest volume of ballast water was discharged at New Plymouth during 1999–2007, with the maximum of 1,151,342.62 tonnes in 2002. The total vessel discharge counts, and DWT also decreased at New Plymouth after this year (Figure 13 and Figure 14). After New Plymouth, the second highest discharge volume was observed at Lyttelton in 2006. Lyttelton received the highest overall ballast discharge between 2004 and 2006. Tauranga also received a high volume of discharged ballast water during 1999–2007. For most years (1999–2007), Tauranga had the highest values for vessel discharge counts (Figure 13) and DWT (Figure 14), as well as the highest number of visits as an arrival port after Auckland (Figure 5).



Figure 14. Total Dead Weight Tonnage (DWT) of vessels visiting discharge ports during 1999–2007.

The total ballast water discharged at each discharge port for every arrival port for years 1999–2007 is shown in Figure 16. The green colour represents the annual ballast water volumes that were discharged at any port except from the arrival port. The yellow colour shows the annual ballast water volumes for vessels that used their arrival ports as discharge ports. “FOREIGN” represent the annual total ballast water related to the ports that are not among the discharge ports of interest. According to Figure 16, New Plymouth had the highest total annual ballast water discharge volumes in 2001 and 2002 (1,085,613 and 1,080,488 tons, respectively) from the vessels that chose this port as the first port of arrival. This agrees with Figure 15, which shows that New Plymouth received the highest total volume of discharge in these years. After New Plymouth, Lyttelton had the highest total annual ballast water discharge volumes of 932,950 and 667,565 tons in 2006 and 2004, respectively. New Plymouth in 1999 and 2000 was the fifth and seventh and Lyttelton in 2002 and 2005 was the sixth and eighth in terms of the volume of total annual ballast water discharge. Taharoa, Whangarei, and Napier were other ports with the highest annual ballast water discharge. The highest values of ballast discharge were related to the vessels that used arrival port as the discharge port (yellow circles) (Figure 16).



Figure 15. Total ballast water discharge volume (tonnes) of vessels visiting discharge ports during 1999–2007.



Figure 16. The total annual ballast water discharge as a function of arrival port and discharge port for the historical raw data. Large ballast discharge volumes are labelled on the graph in tonnes. The green colour represents the annual ballast water volumes that were discharged at any port except for the arrival port. The yellow colour shows the annual ballast water volumes for vessels that used their arrival port as discharge port/ports. The pink circles represent the annual ballast water related to the ports that the discharge port was not defined (NA). FOREIGN refers to the ports that were not among the discharge ports of interest.

## Part 6 – Other techniques employed for model selection

The models were built using Integrated Nested Laplace Approximations (INLA), Generalized Linear Model (GLM), Generalized Additive Models (GAM) with Tweedie family, Least Absolute Shrinkage and Selection Operator (LASSO), Bayesian estimation of GLM models, and random forests. INLA seemed to be the best model selection method because it can accommodate both binary (presence/absence of biofouling) and continuous response variables (biofouling mass) in the model. However, fit statistics were poor. Based on the results (not shown here), GLM and GAM also did not seem to fit the data well.

## Part 7 - Results of the cross-validation for ballast water

Table 5. The results of the cross-validation for Ballast water analysis showing RMSE values for all the combinations of distance variables day and dwt, and predictor variables arrival port and vessel type for K = 1, 3, and 5. The distance variable day was calculated to account for the seasonal variation in ballast water discharged at NZ ports. The best-performing combination was distance variable DWT with data divided by arrival port and vessel type (in bold, below).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Distance variable | Predictor/predictors | K | RMSE |
| 1 | DWT | Arrival port | 1 | 115085 |
| 2 | 3 | 95244.6 |
| 3 | 5 | 91590.3 |
| 4 | DAY | Arrival port | 1 | 118840 |
| 5 | 3 | 118210.9 |
| 6 | 5 | 116890.9 |
| 7 | DWT & DAY | Arrival port | 1 | 121731.9 |
| 8 | 3 | 95591.6 |
| 9 | 5 | 93253.2 |
| 10 | DWT | Vessel type | 1 | 96946.9 |
| 11 | 3 | 108828.9 |
| 12 | 5 | 114179.2 |
| 13 | DAY | Vessel type | 1 | 119765.0 |
| 14 | 3 | 120040.7 |
| 15 | 5 | 120252.5 |
| 16 | DWT & DAY | Vessel type | 1 | 121237.2 |
| 17 | 3 | 105320.6 |
| 18 | 5 | 117625.7 |
| 19 | DWT | Arrival port & Vessel type | 1 | 106158.8 |
| 20 | 3 | 90096.8 |
| 21 | 5 | **88732.8** |
| 22 | DAY | Arrival port &Vessel type | 1 | 101199.7 |
| 23 | 3 | 101685.3 |
| 24 | 5 | 109021.2 |
| 25 | DWT & DAY | Arrival port &Vessel type | 1 | 114251.1 |
| 26 | 3 | 90433.6 |
| 27 | 5 | 88834.9 |

## Part 8 - The validation of KNN method

The validation of KNN method for predicting discharge volume was tested using existing data of years 1997 – 2005 as training dataset and years 2006 – 2007 as validation data. Figure 17 shows the distribution of predicted total discharge by port and year within the validation data, as estimated by bootstrapping. The vertical black lines in each panel show the lower and upper 5% and 95% quantiles of the empirical distribution respectively, whilst the vertical dotted lines show the true total discharge. It is clear that the KNN algorithm predicts the true volume with different quality in different ports. Figure 18 compares the true total discharge by port with the mean predicted discharge by port, as estimated by the bootstrap. The vertical lines show the mean plus/minus two standard deviations, and the dotted line shows what would be perfect prediction.



Figure 17. Predicted total discharge per port distribution in the validation data. The vertical black lines in each panel show the lower and upper 5% and 95% quantiles of the empirical distribution respectively, whilst the vertical dotted lines show the true total discharge.



Figure 18. Predicted total discharge per port distribution in the validation data. The vertical lines show the mean (dot), plus/minus two standard deviations, and the dotted line shows what would be perfect prediction.