Temporal Analysis and Gravity-Informed Marine Traffic Forecasting for Non-Indigenous Species Risk Assessment Through Ballast Water

by

© **Ruixin Song**

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Computer Science

Memorial University

Supervisor: Amilcar Soares

May 2024

St. John's, Newfoundland and Labrador, Canada

Abstract

Non-indigenous species (NIS) spreading through ballast water and establishing themselves in the new environment threaten biodiversity and marine ecosystems. Ballast water risk assessment (BWRA) models estimate the risk for NIS introduction by ballast water, and the environmental similarity between water source and destination locations is important in these models. Previous BWRA models rely on annual-scale environmental data and potentially neglect seasonal variability in the environmental factors. This research investigates the impact of incorporating monthly-scale environmental data on the evaluation of environmental similarity between source and recipient locations. The statistical comparison reveals that using monthly-scale data generally results in smaller environmental distances across all regions, indicating a higher risk of NIS invasions into Canadian waters than previously estimated with annual data. In addition, this work introduces a novel physics-inspired framework to forecast maritime shipping traffic, enhancing the assessment of NIS spread through global transportation networks. Integrating graph analysis, the gravity model, and the self-attention mechanism from the Transformers, this framework outperforms existing methods, achieving an 89% accuracy for discriminating existing and non-existing shipping trajectories and an 84.8% accuracy in estimating the number of vessels flowing between port areas. This represents more than 10% improvement over the traditional deep-gravity model and nearly 50% improvement over the machine learning regressional models, offering a more accurate tool to identify high-risk invasion pathways and prioritize ballast water management in the future.

Acknowledgements

I would like to extend my deep gratitude to my supervisor, Dr. Amilcar Soares, for his invaluable support and guidance in and beyond this research work.

I am very grateful to the senior researchers Dr. Sarah Bailey and Dr. Gabriel Spadon, who have enriched my understanding and practice of research methods. I am very fortunate to have had the opportunity to collaborate with them in this program.

I would also like to express my sincere thanks to the examiners of this thesis, whose insights and feedback have contributed to the refinement and improvement of this research work.

Additionally, I wish to express my appreciation to all the professors and colleagues I have met in classes and the lab at Memorial University. I have gained a lot of knowledge from the courses and discussions in which they have lectured and participated.

Lastly and very importantly, I want to thank my family and friends for their unwavering love and support. A special to my dear uncle, Zhongyi, whom I lost forever in 2022 but memory and influence remain with me always.

Table of contents

List of figures

List of tables

List of abbreviations

AIS	Automatic Identification System
NIS	Non-indigenous Species
BWRA	Ballast Water Risk Assessment
BWM	Ballast Water Management
ISO	International Organization for Standardization
SDM	Species Distribution Model
OD	Origin-Destination
CPC	Common Part of Commuters
NRSME	Normalized Root Mean Square Error

Chapter 1

Introduction

1.1 Background

Invasion of non-indigenous species (NIS) is the process of species naturalizing themselves in non-native regions upon introduction by human activities [2]. During the last decades, records of NIS invasions have seen significant growth due to the increase in globalization [3, 4]. Specifically, increased shipping and other human activities across geographical regions have brought the invasion issue to center stage, threatening biodiversity and ecological balance [5, 6]. Ballast water is used to keep vessels in balance during travels. According to studies on the risk of invasion of NIS, the intake and discharge of ballast water is one of the main pathways for the spread of NIS across aquatic ecosystems [7]. To tackle the challenge of invasive species, various approaches have been developed to assess the risk posed by NIS transported via ballast water over the last decades [8, 9, 10, 11, 12, 13, 14, 15]. This research field is commonly known as ballast water risk assessment (BWRA).

In the BWRA studies, ballast water reports submitted by individual ships are an essential source of data documenting the time and locations of ballast water intakes and discharges. This information can be used to model the NIS trips and evaluate the level of invasion risk, which is crucial to ballast water risk management. While many studies have focused on the impact of regional dissimilarity in environmental factors on the survival of invasive species, seasonal variations in environmental factors have rarely been addressed and implemented in risk assessment applications. This has left a critical topic for further research and tool development.

Additionally, acquiring ballast reports can be difficult when BWRA studies are conducted globally due to various ballast water policies of countries and regions with limited geographic areas to conduct BWRA studies. Therefore, there is a demand for alternative data sources that can provide information on ballast water origins and destinations. The Automatic Identification System (AIS) is equipped on each ship to share the ship's real-time locations, and therefore, AIS data contains the ships' mobility information and is popular in the research of marine domains [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28]. As the ballast water is closely associated with ship movement, AIS is a promising data source for future studies of NIS introduced by ballast water.

1.2 Problem Formulation

The environmental similarity is an essential measure in risk assessment of NIS spreading through the ballast water. More specifically, it measures the similarity between environmental factors of locations, including temperature and salinity on the sea surface [29]. Previous studies have evidenced the seasonal and month-to-month variation of sea temperature and salinity [30, 31, 32, 33]. However, this variation has rarely been considered in existing BWRA studies [11, 13, 15] and the application [34]. A previous study that models the risk of invasion through ballast water has considered the variation of risk by seasonality [12]. Given the potential impact of these seasonal variations in risk assessment, conducting the risk assessment with monthly environmental data is crucial, and can provide insight into the risks involved and allow comparison with previous BWRA models that used annual environmental data. In this way, we can better understand the impact of seasonal variations on BWRA by quantifying differences in risk levels.

As maritime shipping traffic can provide BWRA with a more detailed evaluation, shipping data is also included in BWRA studies to represent trajectory information [11, 12] and to construct the shipping network [15]. In addition to risk levels evaluated between locations with environmental conditions, shipping intensity can affect the cumulative risks on the shipping routes [12]. Therefore, predicting shipping traffic conditions can help to understand future ballast water risk assessments, as the distribution of risk can be predicted, allowing interventions from ballast water management. The shipping traffic prediction can be considered an origin-destination (OD) flows modeling problem, and in OD mobility studies, the gravity model based on the gravity-law principle has been prevalent for many years and across the field of studies including human mobility [35], urban traffic [36], economic trading [37], epidemic spreading [38, 39] and container shipping [40, 41]. The traditional gravity model values the “population” of two locations and the distance between them, and recent studies have explored additional relevant features associated with the flow prediction with machine learning models [42] and deep neural networks [43]. These studies incorporated the Gravity model with machine learning and deep learning techniques, showing the effectiveness of these models in solving practical problems. This has motivated us to model a physics-inspired deep learning framework for predicting ship traffic flow. In our study, in addition to the ship fluxes as “population” and the distance for gravity feature, we employed global economic trade data [44] and graph features analyzed from the global shipping network to extend the number of attributes. Also, when modeling this problem, we utilized Transformer architecture [45] as we hypothesized that the self-attention mechanism could help to capture more intrinsic patterns among features and the ship traffic flow prediction.

In summary, the problems we aim to solve in this work are formulated as the following Research Questions (RQ):

- **RQ1.** To what extent do seasonal variations in environmental factors influence the outcomes of BWRA models? Specifically, how do these variations impact the assessment of NIS risk?
- **RQ2.** Can the integration of the gravity model, graph analysis, and modern deep learning framework enhance the prediction of future ship flows? How do the predicted flows provide insights into shipping intensity on routes with different NIS risk levels?

1.3 Contributions

As stated in Section 1.2, the environmental-based risk assessment is affected by both the evaluated risk levels (i.e., the computed environmental distances) and the shipping intensity. Therefore, we aim to explore the ballast water risk assessment from two different perspectives in the thesis: (i) to analyze the impact of environmental factors' temporal variability posed on the BWRA and (ii) to model a framework that forecasts future maritime ship traffic flows to reflect the shipping intensity, thus helping to predict risk more accurately.

In (i), we explore the impact of the temporal variability of environmental factors and contribute to the following:

- We examined the temporal variation of sea-surface temperature and salinity at ports worldwide by matching each port with the nearest monthly environmental data throughout the year. In this way, ports assigned with monthly temperature and salinity values can show the interannual variation of environmental factors.
- We used a BWRA tool based on the evaluation of environmental dissimilarities and quantified the influence of environmental variation on the calculations of

environmental distances by statistical analysis. These evaluations were based on a case study focused on ballast water discharges in Canadian waters. Statistical analysis shows the difference between the risk evaluated by the monthly and annual environmental variables, and this difference is more significant in the Atlantic and Arctic discharge areas, showing the impact of temporal variability of the environmental variables in the risk assessment.

- Since ballast water is sourced from different geographical regions around the world to Canadian ports, we further explored the temporal variation of BWRA across unique source-destination port pairs and targeted port pairs with more risk variation throughout the year. This analysis provided insights into the evaluated risk variation affected by both the intake time and the source locations of ballast water and offered reference for future ballast water risk management and tool development.

In (ii), we propose a physics-inspired deep learning framework to predict vessel traffic flows and thus provide insights into the ballast water risk assessment. The main points we have contributed in this study include:

- We built a global shipping network from 2017 - 2019 AIS data and analyzed the graph features. Since disconnected components were detected, a link prediction task was performed to target trajectories that are most likely to have shipping activities. This action provides prior knowledge for the subsequent gravity-informed deep framework.

- We proposed a deep framework for predicting ship traffic flow and named it *Transformer Gravity*. It is a combination of stacked Transformer layers and gravity model features. It covers critical gravitational components like shipping flux density, port-to-port distance, and additional factors such as international trade volumes and port centrality measures. Our *Transformer Gravity* model surpasses the traditional Deep Gravity model by over 10% and outperforms other machine learning regression models by nearly 50%, all while maintaining low variance and high reliability in its predictions.

- We calculated environmental distances for all possible pairs of ports as a measure of ballast water risk. Subsequently, we depicted the distribution of these distances in relation to predicted shipping traffic between the ports. This prediction closely aligns with the environmental distances observed in actual shipping traffic. This alignment highlights the effectiveness of our *Transformer Gravity* model in forecasting marine ship traffic and informing BWRA. Furthermore, it highlights the model's potential for future OD mobility studies in broader fields.

1.4 Thesis Outline

This is a thesis by articles and is structured in the following chapters:

Chapter 2 presents the literature that covers essential topics in this thesis, including ballast water risk assessment studies in the past decade and their connection with shipping activities, literature on shipping network analysis and related graph metrics, and physics-inspired OD models for mobility studies.

In Chapter 3, we investigate how the temporal variability of environmental factors influences the results of BWRA. Additionally, we analyze the discharge of ballast water from multiple global regions into Canadian waters. This analysis aims to determine which regions and specific intake times pose a more significant threat to Canadian aquatic ecosystems. This work entitled "A temporal assessment of risk of non-indigenous species introduction by ballast water to Canadian coastal waters based on environmental similarity" has been published in the journal *Biological Invasions* [46]. [1]

In Chapter 4, we propose a comprehensive pipeline for maritime shipping traffic

[1] Reproduced with permission from *Springer Nature*.

flow prediction, consisting of stacked transformer layers and gravity-inspired models. Then, the predicted shipping flows represent the upcoming intensity of the shipping activities and thus can inform the future BWRA. This work entitled "Gravity-Informed Deep Learning Framework for Predicting Ship Traffic Flow and Invasion Risk of Non-Indigenous Species via Ballast Water Discharge" is under review in the journal *Scientific Reports* [47].

Finally, in Chapter 5, we summarize the research findings presented in this thesis. We also state the ongoing challenges in the related fields that are not addressed in our studies and suggest potential future research topics to be explored further.

1.5 Co-authorship Statement

This is a thesis by articles that includes collaboration with other researchers.

Chapter 3 appears as the published work in the journal *Biological Invasions*, coauthored with Dr. Amilcar Soares, Dr. Sarah A. Bailey, and Yashar Tavakoli. In this work, Amilcar Soares and Sarah A. Bailey conceived and supervised the project. Amilcar Soares and Ruixin Song designed the methodology. Ruixin Song collected and analyzed the data and led the writing of the manuscript. Yashar Tavakoli assisted with statistical analyses. Sarah A. Bailey and Amilcar Soares enhanced and facilitated the manuscript.

Chapter 4 is the last revised version submitted to the journal *Scientific Reports* and is currently under review. This work is coauthored with Dr. Gabriel Spadon, Dr. Amilcar Soares, Dr. Ronald Pelot and Dr. Stan Matwin. Ruixin Song and Gabriel Spadon conceived the idea and designed the methodology together with Amilcar Soares. Ruixin Song and Gabriel Spadon performed the formal analysis and data modeling and prepared the original draft. Ruixin Song, Gabriel Spadon, and Amilcar Soares reviewed and edited the manuscript. The project is supervised and funded by

Amilcar Soares, Ronald Pelot, and Stan Matwin. We also thank Dr. Sarah A. Bailey for providing valuable insights on this manuscript.

Chapter 2

Literature Review

2.1 Ballast Water Risk Assessment for Non-indigenous Species

Aquatic non-indigenous species (NIS) can damage biodiversity and threaten marine ecosystems, and this invasion issue has drawn attention during the last decades. Since ballast water is one of the most important pathways to introduce NIS [48, 7], multiple studies have conducted evaluations of NIS introducing risk by ballast water discharges. According to a ballast water risk assessment survey [29], environmental similarity matching and species-specific are primary methods for assessing the invasion risk of NIS by ballast water discharges. While the species-specific approach requires a large amount of up-to-date species data to support the analysis, the environmental-based method focuses more on the dissimilarity of environmental conditions of the source and destinations.

Regarding the environmental variables used in the environmental method, salinity, and temperature are the most important environmental factors for the survival of aquatic species [49]. A study used the different environmental variables and analyzed the significance of these factors correlated with the real invasion conditions by testing on hundreds of species [50]. The results validated that fewer, more relevant variables and Euclidean distance calculations can bring environmental similarity assessments closer to true invasion conditions. This way of building environmental vectors has been adopted in the more recent study of developing a risk evaluation tool for ballast water discharges in Canadian waters and is used by Transport Canada [34], which used ballast water reports submitted by ships to identify the ballast water source and the discharged water volume.

Table 2.1: Comparison of risk assessment models with network employed

Model	Method summary	Dataset coverage	Analysis range	Network-related work
Risk assessment for ships destining to Laurentian Great Lakes [11]	Matching environmental conditions between ports in a global range and stepping ports with higher order connections	Port locations, shipping voyages, environmental data	Global, a local case study at Laurentian Great Lakes	Considering higher-order connections (up to 5 steps) in shipping activities
Probability model for invasive species risks prediction [12]	Combining the whole shipping routes with the statistical submodels for comprehensive risk prediction	AIS data, ballast report, environmental data, invasion events	Global	Visualizing high invasion probability shipping routes
Risk model for predicting the spread of aquatic species [14]	Adopted from the previous work (H. Seebens et al., 2013) with model validation	Ship movement, ballast report, environmental data, invasion events	Global	Considering number of voyages as shipping intensities
Higher order network with risk model implemented for NIS risk assessment [15]	Introducing the higher-order network (HON) to the NIS risk probability model. Comparing the network with the first-order network (FON) model by network analysis	Ship movement, ballast report, environmental data, biogeographical regions, NIS distribution	Global	Performing network analysis including betweenness centrality, density and clustering coefficient to compare HON and FON risk models

Furthermore, many BWRA studies used shipping information to assist in the assessment of NIS invasion risk. These include the risk assessment for ships that discharge ballast water in the Laurentian Great Lakes and analyzed the higher order in the shipping network [11], and the probabilistic model for global BWRA that informed the variation of invasion risk across different bioregions and identify the high-risk routes [12]. In addition to assessing the risk of ballast water, a more recent work analyzed the risk of biofouling using the probability model of [12] and considered higher orders in the shipping network [15]. The introduction stress in the submodel was modified to adopt species accumulation and survival factors for biofouling risk evaluation. More information about these NIS risk assessment models is listed and summarized in Table 2.1.

Our study [46], as presented in Chapter 3, leverages the decision support tool utilized by Transport Canada [34]. This tool is based on the high-impact, environmental similarity-driven BWRA model proposed by Keller et al. [11]. Additionally, we are inspired by the work of Seebens et al. [12], which examines risk variation in ballast water discharge ecoregions but only reflects trends of risk variation. Our method extends this point by quantifying the impact of temporal variability in environmental factors on BWRA outcomes. We employ the Wilcoxon signed-rank test [51] to analyze the BWRA results using monthly and annual environmental data. This analysis demonstrates significant differences in pairwise environmental distances evaluated from monthly- and annual-scale environmental data. For discharging areas in Canadian waters, it shows that using traditional annual-scale data has underestimated the NIS invasion risk. Our further analysis of the risk variation among ballast water source-destination port pairs has revealed the regions and specific times of year that significantly contribute to the high-risk invasion of ballast water into Canada. These findings can inform the ballast water management from both spatial and temporal perspectives, contributing to the related BWRA studies.

2.2 Shipping network and Graph Analysis

Although many BWRA studies utilized ballast water reports submitted by ships as the data source, shipping data is still essential for BWRA and especially meaningful while the acquisition of ballast reports can be limited due to various ballast management policies across countries and regions. In addition to being closely linked to ballast water transportation, analysis of shipping networks comprising shipping data can provide insights into maritime transportation patterns for the NIS traveling with ballast water. Specifically, the Automatic Identification System (AIS) equipped on each ship reports the real-time location and the condition of voyages [52] and has become a reliable shipping data source for shipping network construction. Saebi et al. [15] have built a higher-order risk network from shipping routes and port pair invasion risk and then calculated a series of network features, including the betweenness centrality, the clustering coefficient, and the connected components.

However, such analysis of the shipping network for NIS risk is rarely seen in previous studies, and network analysis has been used in other maritime application scenarios involving local network analysis for European ports [53], shipping network evolution patterns worldwide [54], evaluation of centrality variation during years to understand the variation of global shipping conditions [55], container shipping network analysis with vectorized network representation [56], and also a comprehensive review of the complex network analysis with maritime shipping traffic [57]. In these studies, the shipping networks are represented by the real structure: they use nodes (vertices) to denote ports, and links (edges) to represent shipping routes[1], number of voyages or betweenness of risk values.

As several studies mentioned above have incorporated graph metrics in their shipping network analysis, in the following part, we discuss several essential concepts in graph analysis and their application in the shipping network: (1) betweenness centrality, (2) closeness centrality, and (3) page rank.

(1) Betweenness centrality: betweenness centrality [59] measures how often a node appears on the shortest paths between other nodes in a graph. In the appendix of Chapter 4, Equation 4.16 shows the detail of betweenness centrality calculation. In shipping networks, ports with high betweenness centrality are often critical. They are usually regarded as transit points that carry a significant amount of traffic between various ports on the network.

(2) Closeness centrality: closeness centrality [60, 61] sums the shortest distances of one

[1] To avoid confusion: usually ”nodes and links” are from a network perspective, while ”vertices and edges” are the terms used in graphs [58]

node to all other nodes in a graph and then calculates the reciprocal to measure how "central" a node is in this graph. Equation 4.14 shows the calculation of the closeness centrality. In shipping networks, ports showing high centrality values are usually those located in more central positions and generally have more direct connections with other ports [62].

(3) PageRank: PageRank [63] was first proposed for ranking the web pages in the search engine results. The core idea of PageRank is to measure the rank of a page based on the ranks of the pages linked to this page. In a page graph, each page is represented as a node and is initially assigned the same rank, and then the algorithm traverses the nodes and adjusts the ranks by iterations. A link from a highly ranked page to another page is considered a strong endorsement and thus increases the rank of the latter page. In the shipping network, applying PageRank calculation to shipping ports can help to identify the important hubs that have a large amount of incoming and outgoing shipping connections and are more likely to have stronger connections with other shipping ports.

Motivated by these studies of graph analysis on shipping traffic networks [64, 53, 56] and on NIS risk network [15], our work [47] presented in Chapter 4 models the shipping data as a directed and weighted graph, leveraging the investigated graph metrics, including betweenness centrality, closeness centrality, and PageRank, to enrich the feature set to inform the prediction of shipping traffic flows. Even though many studies have conducted graph analyses on shipping networks before, we calculate and value these graph measures as they can show the centrality and importance of a shipping port in the local and global network. These metrics outbound the single origin-destination mobility prediction mode and involve more information from the whole network to inform the forecast of shipping traffic flows.

2.3 Physical-inspired origin-destination mobility

The mobility study covers several topics, including trajectory prediction, mobility flow prediction, and next-location prediction [65]. Since our *2nd* research question aims to forecast the intensity of shipping traffic in the global shipping network to support BWRA, we focus on "mobility flow prediction" in this section. Our investigation starts with the gravity model and other physics-inspired models for origin-destination (OD) mobility flow prediction. Then, we discuss the limitations of these traditional physics models in the prediction task and explore novel physics-inspired mobility models incorporated with machine learning and deep learning techniques.

2.3.1 Gravity model and other physics-inspired models

The gravity model can be derived from Newton's law of universal gravitation [66], which describes that the force between two objects is proportional to their mass and inversely proportional to the distance between them. In the 1940s, Zipf suggested a similar idea for understanding human movements, which proposed that the number of people traveling between two places could be estimated by looking at the populations of these places and their distance apart. In detail, the movement of individuals T from one community to another using a simple formula: the product of the two communities' populations P_i, P_j divided by the distance D between them [35]:

$$T = \frac{P_i P_j}{D} \tag{2.1}$$

This early study set the first example of using the gravity model for human mobility flow prediction. Later, the gravity model has been widely applied to studies in multiple disciplines, including traffic pattern [36], economic interaction [37], mobile phone communication [67], and cargo shipping [41, 40]. Gravity models in these studies have introduced parameters. Taking human mobility in Equation 2.1 as an example, these parameters include powers to the populations P_i and P_j and a deterrence function on the distance D. This enables the model to adjust parameters based on historical data to give better predictions.

Despite the prevalence of gravity models in solving mobility prediction problems in many research areas, some limitations should not be ignored. First, the assumption that mobility flows are associated with "masses" and "distances" in the gravity model can be over simplistic. In practice, more factors can affect mobility flows, such as urban infrastructure and job opportunities for human mobility [42], economic conditions of countries, and bilateral trade for container shipping. Second, the parameter settings in the gravity model lack rigorous theoretical basis [68], leading to deficiencies such as the power law parameters on the source-destination "masses" hard to define, and parameters are highly dependent on empirical data that the model cannot be migrated to predict flows in regions without historical information feed. Also, the gravity model considers only the features related to the origin-destination pairs but overlooks the flow dynamics in the network. Features of other nodes and edges may also affect flow prediction between specific pairs. This point is merely mentioned in related studies is one point we aim to explore in this thesis.

In recent years, alongside the traditional gravity model, other physics-inspired models have emerged to explain spatial interactions. Among these, the radiation model for human mobility was introduced by Simini et al [68] and later applied to study patterns in the urban commuter network [69]. This model draws inspiration from the radiation absorption process in physics, incorporating the population size and the number of job opportunities within a defined radius. Unlike the gravity model relying on distance and population mass, the radiation model suggests that job opportunities more directly influence human mobility. Moreover, its parameter-free nature addresses the shortcomings of the gravity model, where the parameters often lack theoretical bases. In the context of human mobility, a recent study represents commuter flows in terms of field vectors and then compares the performance of gravity and radiation models for flow prediction over the established vector fields [70]. This integration of field theory and OD models provides insights into urban mobility and future studies in relative areas.

2.3.2 Mobility with machine learning and deep learning

Recent advances in machine learning and deep learning techniques have brought new opportunities to traditional physics-inspired models. A recent study leveraged multiple urban indicators with features derived from the gravity model and used machine learning algorithms to predict human mobility within the urban commuter network [42]. Compared to traditional gravity models, using machine learning and the enriched feature set has improved predictive accuracy and also addresses the challenge of defining multiple parameters that are traditionally difficult in gravity models. Further, in the research of human mobility, Simini et al. introduced a deep learning framework, DeepGravity, which integrates the gravity model with urban features to analyze inter-city individuals' movement [43]. This framework innovatively interprets the gravity model as a multi-classification task, showing better performance over traditional methods when incorporated with the same features. Additionally, a study on urban traffic utilized the graph convolutional networks (GCN) to estimate the urban taxi mobility [71], which valued the graph structure of mobility flows and enhanced the feature vector representation on estimating urban taxi mobility.

The studies referenced above are important in shaping our analysis of shipping traffic mobility. Our research draws inspiration from these works by integrating comprehensive feature sets directly relevant to real-world challenges. Also, we leverage modern deep-learning frameworks to deal with the complexity of mobility patterns. Further, our adoption of graph representation and analysis of the mobility network enables us to capture graph metrics from the network to enhance the flow prediction. All these elements together have inspired our methodology and findings in this work.

Chapter 3

A temporal assessment of risk of non-indigenous species introduction by ballast water to Canadian coastal waters based on environmental similarity

3.1 Introduction

The biological invasion process can be divided into several stages, including transport, introduction, establishment, and spread [72], all of which must be successfully passed for a non-indigenous species (NIS) to be considered 'invasive', which refers to species introduced into new habitats with significant detrimental impacts on native organisms [73]. Overcoming the barriers associated with each stage depends on multiple factors involving propagule pressure [74, 75], environmental similarity [29] and species' traits [76]. The multiple stages and interacting factors are challenging for risk assessment and proactive management, especially in aquatic ecosystems where physical access is limited and data/information are incomplete, unvalidated or not standardized across regions [77, 78]. Increased human activities across biogeographic regions have brought the issue of biological invasions to the forefront, with the main vectors for aquatic NIS introduction and spread being ballast water discharge and biofouling on ships [7, 48].

Ballast water has been responsible for the transport and introduction of a variety of aquatic species across many regions, including bacteria, fungi, plants, and animals [79, 80, 48]. Therefore, a series of ballast water risk assessment (BWRA) tools were launched during the last few decades to guide management activities based on three main approaches: environmental matching, species' biogeograpy and species-specific [81]. The species-based methods call for a multitude of data such as species' geographic distribution, life cycle attributes, and physiological tolerances to assess the potential for introduction and establishment in a new environment [29]. A key issue with species-based approaches is that, by definition, species fundamental niches differ from their realized ones, so there is a need to continually update data based on emerging information with each new species location record. In contrast, the environmental matching strategy is a more general approach, with more readily-available data that do not need such frequent updating. However, given climate change and cyclical climate variability, environmental data should be updated periodically to maintain the validity of the analysis. Furthermore, the environmental matching approach enables the customization of the environmental variables according to the needs of the assessment.

Early BWRA models using the environmental matching approach include a risk assessment in Nordic coastal waters based on salinity and climate factors [8], and the GloBallast BWRA [9, 10] led by the International Maritime Organization (IMO) which contains more than 30 environmental parameters. More recent examples include an assessment of salinity match between donor and recipient ports for ships traveling between canals and oceans [13] and probabilistic models integrating the environmental matching method to assess the risk of NIS invasion through ballast water [12, 14, 15]. Temperature and salinity are consistently included in these models as environmental matching variables, since they are considered the most critical factors contributing to the survival and establishment of aquatic species [29, 11].

A recent study comparing multiple sets of environmental variables against species distribution data shows that models using fewer, but more relevant, variables can perform better than those including many variables in the environmental matching approach [50]. As the most important factors for BWRA, sea surface temperature and salinity can be expressed with different measurements (e.g., minimum or maximum annual temperature [49], annual average salinity, etc.). Since the publication of an influential model using annual average environmental data [11], many subsequent studies have conducted environmental matching assessments following the same variable set, including the BWRA tool currently used by Transport Canada [34, 82] which is the baseline model in this study. Although prior models are stable and provide insight on the likelihood of NIS introduction, the use of annual averages of temperature and salinity has limitations primarily related to insensitivity to seasonal variability, which potentially affects the probability of any introduced NIS survival and establishment at a given time point.

Therefore, this study uses monthly temperature and salinity values obtained from the World Ocean Atlas 2018 [83, 84] within the baseline model to explore the impact of temporal variation in environmental factors on BWRA, based on a case study of ships destined for Canadian ports in 2019 and 2020. The outcomes of the monthly and annual temporal scales are compared statistically with the null hypothesis that there is no difference in risk estimates using monthly vs. annual values in the calculation of environmental distance between ports. Due to wider seasonal variation in sea surface temperature in temperate climate zones, we hypothesize that ballast water from temperate ports of origin will show greater variability in environmental distance calculations for monthly vs. annual scale assessments. In addition, the opposite seasons in the northern and southern hemispheres may result in Canadian ports with low overall temperatures being at higher risk of NIS survival and establishment from ballast water originating from the southern hemisphere winter. To explore these hypotheses, we calculated risk values for each pair of ports during different months of the year using fixed time intervals and explored the interannual risk variability.

3.2 Methods

3.2.1 Fundamentals of the baseline risk model

The baseline model used in this study [34] is the practical tool used by Transport Canada for assessing ballast water risk as an essential input into decisions concerning derogation requests and contingency measures [82]. Canada requires ships to submit ballast water reporting forms, declaring the source port of any ballast water to be discharged in Canadian waters, as well as details about any management activities undertaken (e.g. ballast water exchange and/or ballast water treatment). The baseline BWRA model assesses the risk of each ballast water tank discharge by comparing environmental similarities between source and recipient ports [34]. The model first normalizes the environmental data with a z-score procedure applied to four environmental variables: (i) maximum, (ii) minimum, (iii) average temperature and (iv) average salinity.

More formally, the environmental vectors V are:

$$V = \langle T_{max}, T_{min}, T_{avg}, S \rangle \tag{3.1}$$

where $T_{max}, T_{min}, T_{avg}$ are the normalized maximum, minimum and average temperature, and S is the normalized average salinity of a source or destination location. After, the Euclidean distance is calculated between the four variables for ballast water source (v_s) and destination (v_d) (v_s, $v_d \subset V$) as follows:

$$env\_distance(v_{s_i}, v_{d_i}) = \sqrt{\sum_{i=1}^{|V|}(v_{s_i} - v_{d_i})^2} \tag{3.2}$$

Ballast water management actions that could alter the environmental variables, such as offshore ballast water exchange, are not considered in the model since the tool is used as part of a precautionary management approach considering the 'worst-case' scenario. The assessment can easily be re-executed using geographical coordinates of ballast water exchange as the source location when desired.

As previously described, the baseline model currently uses annual-scale environmental data - mean temperature during the warmest month (as the maximum temperature), mean temperature during the coldest month (as the minimum temperature), annual average temperature and annual average salinity, following [11]. Risk categories are then assigned based on the distribution of environmental distances between all pairwise permutations of ports on a global scale. The distribution of distance values is categorized by the percentiles in Table 3.1.

Table 3.1: Percentiles of environmental distance values and corresponding risk categories based on all possible combinations of global port pairs.

Percentile	Distance value d	Category
0 – 20%	$d < 0.787$	very high risk
20% - 40%	$0.787 \leq d < 1.500$	high risk
40% – 60%	$1.500 \leq d < 2.778$	moderate risk
60% – 80%	$2.778 \leq d < 4.020$	low risk
80% - 100%	$d \geq 4.020$	very low risk

Transport Canada personnel can use these categories as part of prioritization to quickly identify ballast tanks that pose greater risk since categorical data are more easily interpreted than the numerical distance values. In this study, however, only numeric distance values are used in the analysis because the data are continuously distributed and have a wider range of values than the categorical results.

3.2.2 Compilation of monthly and annual environmental data

The compilation of the monthly-scale environmental data was conducted using two datasets: (i) a list of 8,392 global shipping ports with positional coordinates (latitude and longitude) [85] and (ii) monthly sea surface temperature and salinity data downloaded from the World Ocean Atlas 2018 (WOA 2018), available from the National Centers for Environmental Information (NCEI) as the average of six decadal means from 1955 to 2017 following systematic data quality control techniques [84, 83]. The environmental variable values were available at a one-degree grid resolution (i.e., points spaced at approximately 111 km) from January to December. The shipping port locations were matched with sea surface environmental variables based on closest geodesic distance. As the distance for some inland ports to the nearest environmental data point was greater than 2 grid cells (greater than 222 km), the analysis was restricted to coastal ports best represented by the data (all ports farther than 2 grid cells from the nearest environmental data point were excluded, e.g., Laurentian Great Lakes' ports). The match procedure constructed 24 intermediate layers covering 12 months' salinity and temperature data. In each layer, sea surface values were missing for 0.5%-3% of the 31,000-33,000 environmental data points. Since the percentage of missing values was relatively small, these points were dropped for each layer and the closest match procedure was rerun. The layers were then combined to create a dataset of global shipping ports with monthly environmental values.

The standard deviation (STDEV) of the 12 months' environmental values was calculated at each port to examine how the environmental variables change during the year on a monthly basis. The STDEVs at the ports were used to generate a raster layer, and the equal interval method was applied to categorize the values into equal bins.

3.2.3 Evaluation of monthly vs. annual environmental distances for ballast water discharges in Canada

First, we extracted ballast water records (i.e., location and dates when ballast water was taken up and discharged, for individual ballast water tanks) from ballast water reporting forms submitted by ships entering Canadian waters in 2019 and 2020, as stored in the Canadian Ballast Water Information System [82]. The tank records for the two years were processed separately to see if there was a similar/stable pattern across years. Next, we calculated environmental distance values for each pair of ballast water source-recipient locations using annual and monthly environmental data as inputs to the baseline model and created density distribution plots to visualize the difference between the two temporal scales in each year.

We then calculated the difference in environmental distance values produced using the monthly and annual environmental datasets, subtracting the annual distance from the monthly distance for each tank record: $distance\_diff = env\_distance_{month} - env\_distance_{year}$. The resulting difference values were divided into two sets, one with positive difference values and the other with negative difference values. Positive difference values result when a port-pair was at lower risk (had a greater environmental distance) using the monthly environmental data, while negative difference values result when the port-pair was at higher risk (had a lower environmental distance) with the monthly environmental data. These two sets of difference values were examined separately, selecting the 75% and 90% percentiles of positive and negative differences as thresholds of importance, generating four categories:

1. (Positive difference greater than 1.586) - port-pairs with much lower risk using monthly environmental data

2. (Positive difference between 0.835 and 1.586) - port-pairs with lower risk using monthly environmental data

3. (Negative difference between -1.720 and -2.248) - port-pairs with higher risk using monthly environmental data

4. (Negative difference lower than -2.248) - port-pairs with much higher risk using monthly environmental data

Difference values were averaged across all individual ballast tanks discharged at each Canadian recipient port, and those falling within the above categories were marked in darker colors on a map to visualize ports with more pronounced differences in environmental matching at the two temporal scales. Ports with average difference values outside these categories were marked with lighter colors on the map, indicating ports without notable changes in assessed risk after using monthly environmental data.

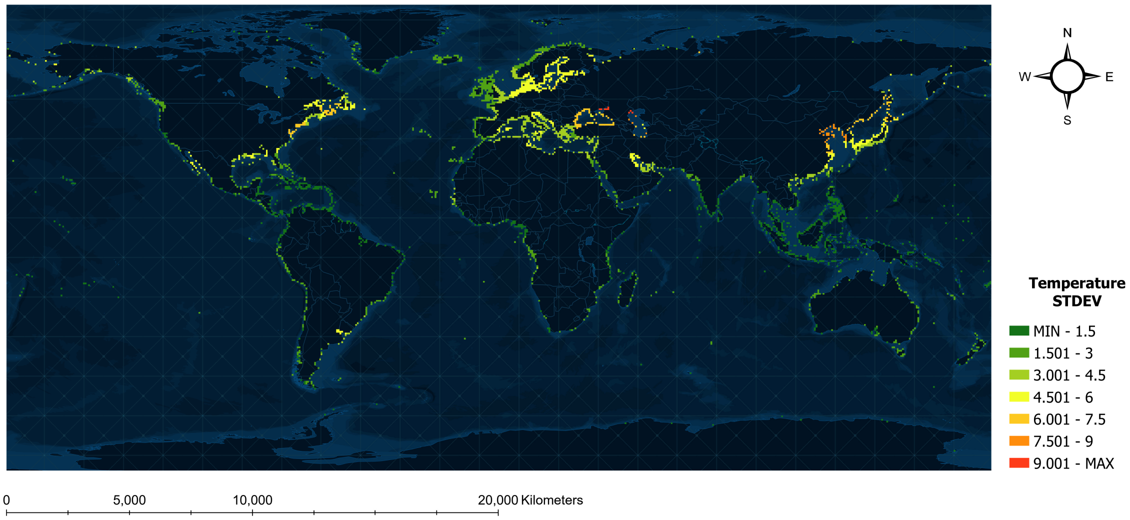
In addition to the categorical assessment of the pronounced differences, statistical tests were conducted to evaluate the significance of the overall difference between the monthly and annual environmental distances. Since the distribution plots showed that the distributions were skewed, non-parametric tests were used. Monthly and annual environmental distances were calculated for each port pair in the ballast water tank data, pairing the monthly distance value (based on actual date of the ship trip) with the annual distance value one to one (i.e., the baseline model was run for each ballast tank source-destination record using both scales of environmental data). The Wilcoxon signed-rank test for paired samples [86] was used to examine whether the differences in the two calculations were statistically significant, with the null hypothesis that the differences between the two samples were symmetric about a real number μ such that the two samples can be recognized as similar distributions. We used the function ”wilcox.test” in R [87] to perform the evaluation with a significance level α of 0.05. The Wilcoxon test effect size, function ”wilcox_effsize” in the *rstatix* package

[88], was used to examine the strength of the differences across all paired samples together and for paired samples aggregated by region (Atlantic, Pacific and Arctic). The statistical tests were performed on the paired environmental distances in 2019 and 2020 separately to verify whether the patterns of differences in environmental distances were stable across these two years.

3.2.4 Standardized analysis of monthly environmental distance variation

Since the statistical analysis conducted in Section 3.2.3 may be biased by specific factors in the Canadian ballast water data such as shipping intensity between specific port pairs or the actual date (month) of different ship trips, a standardized analysis was conducted which excluded replicate tanks and examined differences in monthly vs. annual environmental distances across each unique source and destination port pair across all months in the year (rather than only for the dates of actual ship trips in the Canadian dataset). We calculated the average voyage time (τ) for each unique port pair based on dates reported in the ballast water data.

We then cycled the start date of the voyage from January to December, using τ as a fixed time interval to calculate 12 environmental distance values representing a one-year cycle for each port pair using Eq. 3.2 with the corresponding monthly environmental data. The standard deviation of the 12 environmental distance values was then calculated for each port pair to explore the magnitude of change in environmental distance during one year. Furthermore, the source ports were grouped into regions to explore patterns in environmental distance differences by region across months. The country code and regions used followed the ISO-3166 Standard [89].



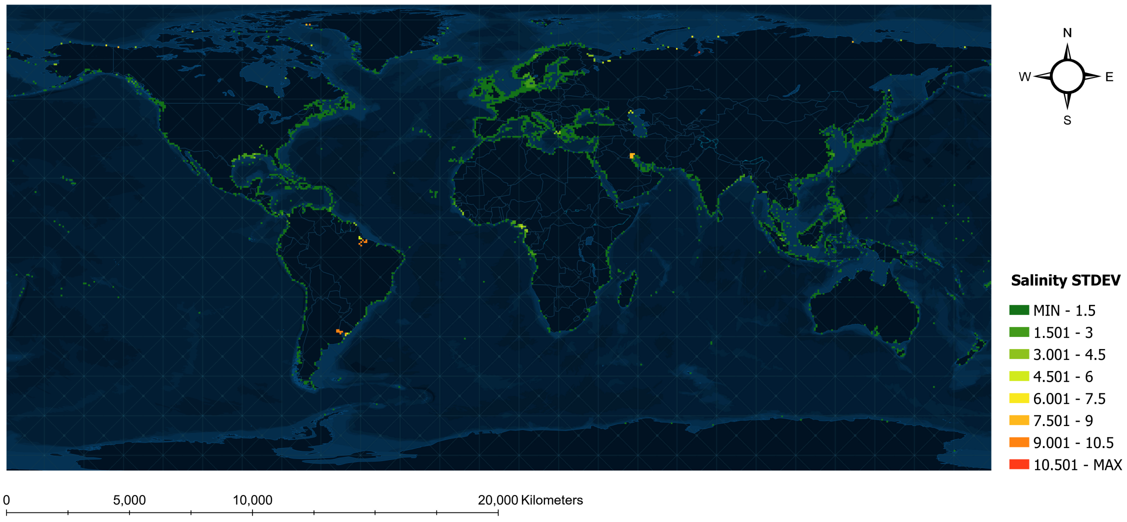
Esri, FAO, NOAA, USGS

Figure 3.1: Temporal change in temperature at global coastal shipping ports, illustrated by standard deviation (STDEV) of monthly decadal average values. STDEVs close to zero (dark green) indicate less change in temperature during a year, while large STDEVs (red) indicate greater temperature change.

3.3 Results

3.3.1 Temporal changes in environmental variables at global ports

The standard deviation of monthly decadal average environmental values at global coastal shipping ports across one year can be seen in Fig.3.1 and Fig. 3.2, for temperature and salinity, respectively. Fig. 3.1 shows that temperature changes greater than three standard deviations occur broadly and are greatest in the northern hemisphere, especially in the temperate climate zone. Fig. 3.2 shows that the largest temporal changes in salinity are mainly concentrated in the estuaries of large rivers (e.g., Amazon and Uruguay rivers in South America, Volga River in Eastern Europe).



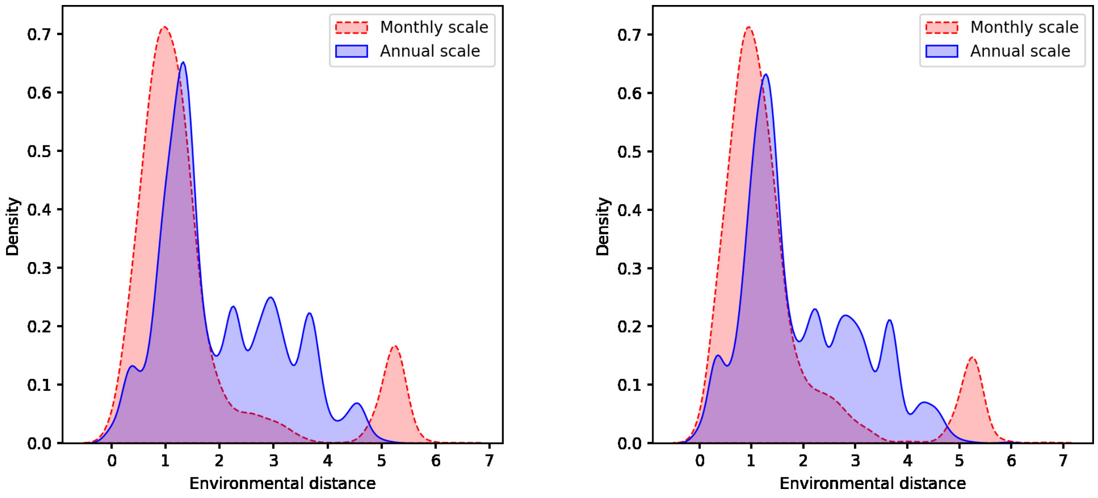
Esri, FAO, NOAA, USGS

Figure 3.2: Temporal change in salinity at global coastal shipping ports illustrated by standard deviation (STDEV) of monthly decadal average values. STDEVs close to zero (dark green) indicate less salinity change during a year, while large STDEVs (red) denote greater salinity change.

3.3.2 Temporal changes in environmental distance across ballast water discharges in Canada

Ballast water source and discharge locations were extracted from 87,951 tank records (7,242 trips) submitted by ships arriving in Canadian waters in 2019 and 2020 [82]. After removing inland ports and discharge locations outside of Canadian water, 51,945 tank records (representing 6,308 ship trips and 1,357 unique source-recipient port pairs) remained for analysis. Figure 3.3 shows the distribution of the environmental distance values produced by the baseline model using annual and monthly environmental data for all ballast tanks discharged in Canadian waters. The distributions of environmental distances in both years show more extreme values when using monthly data (i.e., monthly distributions have more small and large values).

Looking only at the extreme values in the 90% percentile categories, 17,291 tank records are at very high risk, of which 50.03% and 49.96% are destined for the Atlantic and the Pacific regions, respectively. Meanwhile, 71.50% of 7,490 very low-risk tank



(a) 2019

(b) 2020

Figure 3.3: Density distribution plots of environmental distance values calculated using monthly and annual decadal averages for ballast water discharges in Canada in (a) 2019 and (b) 2020.

records were discharged in the Pacific region. Comparing the output of the baseline model using monthly vs. annual environmental data, the proportion of very high risk to very low risk tank discharges increases to nearly 7:3 (monthly) compared to 5:7 (annual).

Fig. 3.4 shows differences in monthly and annual average environmental distances. Positive difference values are records where environmental distances increase (i.e., risk values decrease) after using monthly environmental data. Correspondingly, negative difference values indicate records where the environmental distances decrease (i.e., risk values increase).

Examining these differences spatially, we observed that the cumulative risk across all tank discharges at individual ports can become much higher using monthly environmental data (e.g., Fig. 3.5, in dark red: Kitimat, Port McNeil, Port Alberni, Sechelt, New Westminster on the Pacific coast and Havre St. Pierre, Paspébiac, South Brook, Holyrood on the Atlantic coast). Conversely, the cumulative risk becomes much lower

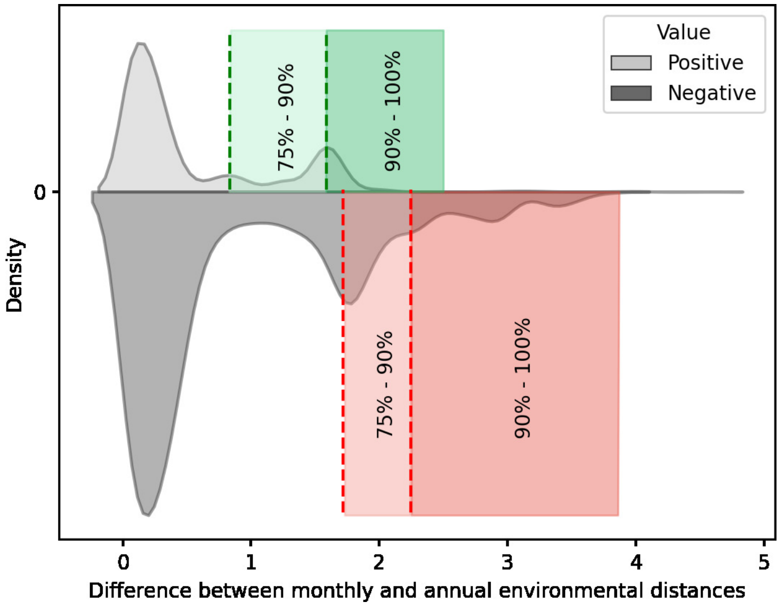
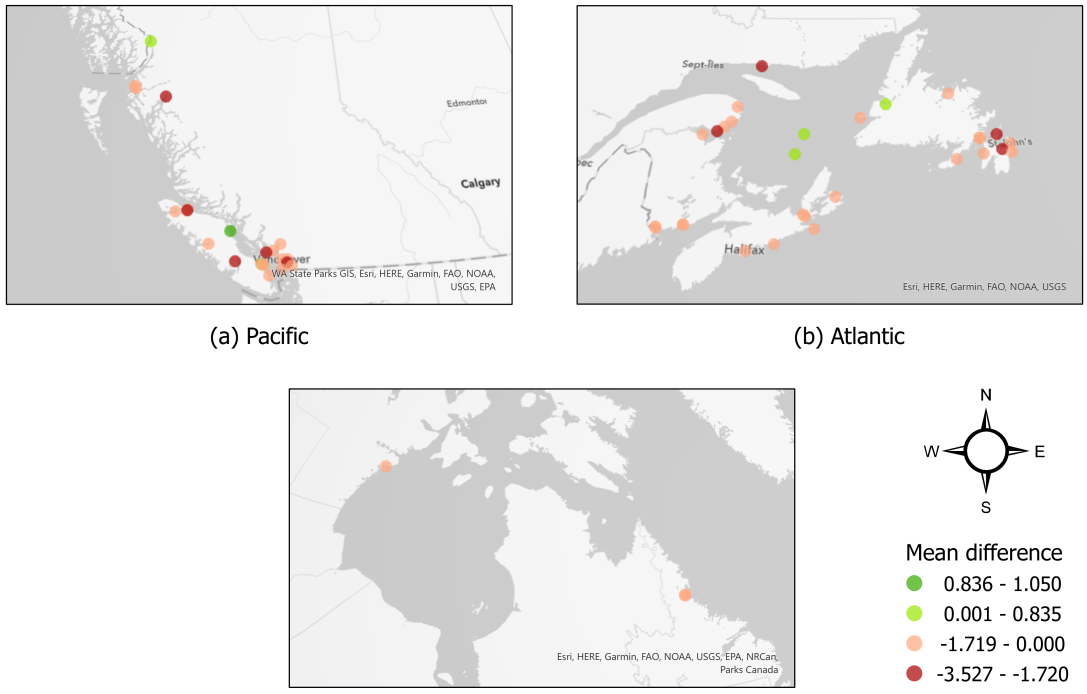


Figure 3.4: Density distribution plot of the differences in environmental distance calculated using monthly and annual decadal average environmental data, with 75% and 90% percentile categories considered as being a significant change marked (dotted lines). Positive values above the zero-axis represent lower risk using monthly environmental data, while negative values below the zero-axis represent higher risk compared to estimates using annual environmental data.


Edmonton
Calgary
Vancouver
WA State Parks GIS, Esri, HERE, Garmin, FAO, NOAA, USGS, EPA
Sept-Îles
St. John's
Halifax
Esri, HERE, Garmin, FAO, NOAA, USGS
Esri, HERE, Garmin, FAO, NOAA, USGS, EPA, NRCan, Parks Canada
N
W
E
S
Mean difference
● 0.836 - 1.050
● 0.001 - 0.835
● -1.719 - 0.000
● -3.527 - -1.720

(a) Pacific

(b) Atlantic

(c) Arctic

Figure 3.5: The average differences between monthly and annual scale environmental distance values at Canadian destination ports. The difference values are attributed to four categories according to the percentile 75% significance thresholds as shown in Fig. 3.4 for the colored areas. Ports marked with orange and dark red have higher risks, while those with light and dark green have lower risks using monthly-scale model.

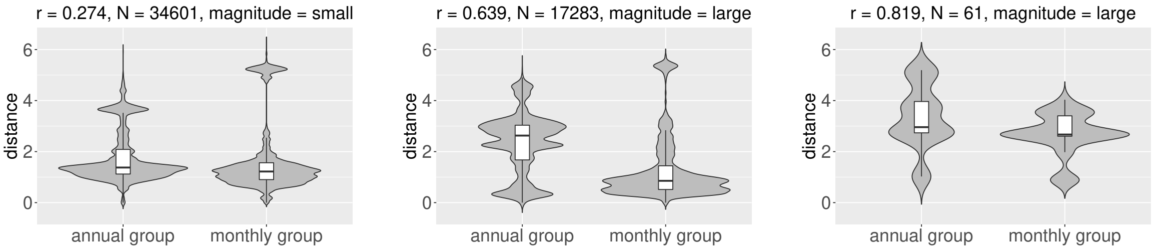
using monthly environmental data for only one individual port (Fig. 3.5, in dark green: Campbell River on the Pacific coast).

Table 3.2 shows the degree of differences between the paired estimates of environmental distance based on monthly and annual data across regions, using the Wilcoxon signed-rank test with effect size r, where the objective of this test is to validate whether there are significant differences between the assessed monthly and annual average environmental distances. The effect size r used to measure the size of difference is largest for the Arctic region, followed by the Atlantic region, while being relatively small for the Pacific region. As the r values for 2019 and 2020 are very similar, the regional patterns in the risk differences are stable across the two years.

Table 3.2: Wilcoxon signed-rank test results comparing environmental distance values based on monthly vs. annual environmental data for ballast tank discharges in Canada during 2019 and 2020. N is the sample size (# of tank records); r is the effect size that quantitatively measures the difference between the paired values, ranging from 0 to 1 where large effect size suggests significant difference. *magnitude* categorizes the effect size as: $< 0.3 =$ "small", $0.3 - 0.5 =$ "moderate", $> 0.6 =$ "large".

	2019			2020		
Region	N	r	magnitude	N	r	magnitude
Pacific	16574	0.282	small	18027	0.267	small
Atlantic	8294	0.637	large	8989	0.640	large
Arctic	45	0.863	large	16	0.845	large

Fig. 3.6 shows the regional distribution of environmental distance values calculated in the baseline model using annual vs. monthly environmental data for the two years of study. Except for some outliers in the Pacific region, the environmental distances based on monthly data generally become smaller in all regions, revealing that when using the monthly data, there is a higher estimated risk of NIS survival and establishment.



(a) Pacific

(b) Atlantic

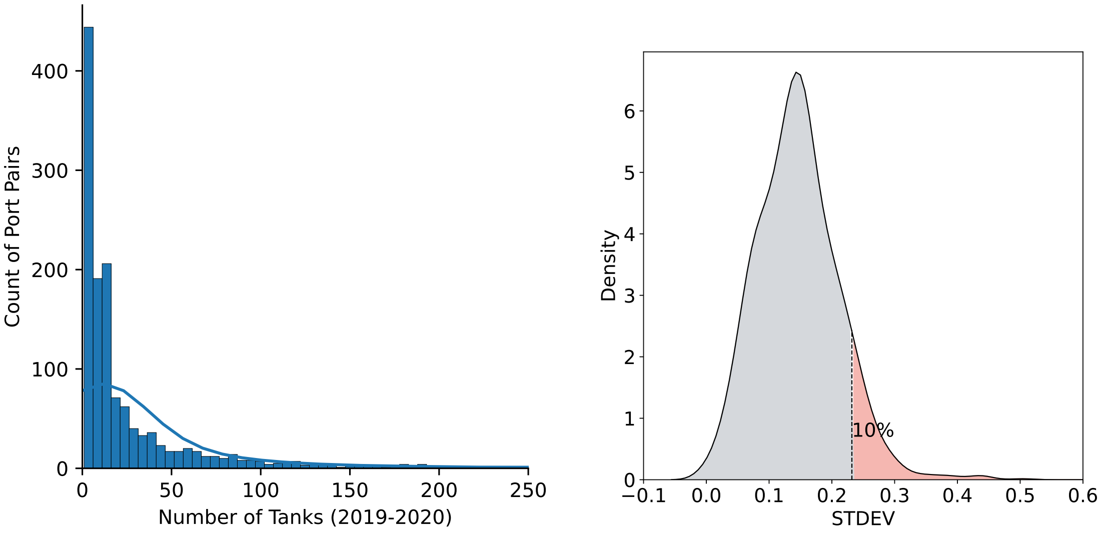
(c) Arctic

Figure 3.6: Violin plots showing the distribution of annual and monthly scale environmental distances, by region (panels a - c = Pacific, Atlantic, Arctic, respectively). The vertical black line show the 1.5 times interquartile range, with white boxes showing the median (center black horizontal line), first and third quartiles (lower and upper box edges, respectively). r is the effect size, N is sample size of the paired distance values, *magnitude* is based on r value.

3.3.3 Temporal variation across unique port pairs

The Canadian ballast water dataset contained 1,357 unique port pairs after the removal of duplicate trips/tanks with the same source-discharge combinations. Most port pairs are connected by only a small number of ballast tank discharges, while a small number of port pairs have a large number of connections (Fig. 3.7a). Port pairs with more than 250 tank connections during the two years are listed in Table 3.4 in Appendix 3.4. The STDEVs of environmental distance values calculated for all unique port pairs across the 12 months of the year is shown in Fig. 3.7b, where large STDEV equates to higher variation in environmental distances during a year. Port pairs having both a large number of ballast tank connections (more than 250 discharges) and high variation in environmental distance during the 12 months of the year (top 10% as shown in the red area of Fig. 3.7b) are presented in Table 3.3. All of these high intensity/high variability port pairs link Eastern Asia to ports located in the Pacific region of Canada.

Further exploration of monthly environmental distance variation by source port region shows how environmental distance can change during the year (Fig. 3.8). The spatial distribution of the 602 source ports across 14 global regions is shown in Figure



(a) (b)

Figure 3.7: (a) Distribution of the number of ballast water tanks (x-asis) transported between port pairs (y-axis). Port pairs with more than 250 tank connections were excluded from the plot (about 2.6%) for visualization purposes (listed in Table 3.3). (b) Distribution of standard deviation of environmental distances for all unique port pairs during the 12-month standardized analysis. The red area denotes the 10% of port pairs with the largest STDEVs.

Table 3.3: Port pairs with high number of ballast tank connections in 2019-2020 and high variation in environmental distance values using monthly scale.

Source Region	Source Port	Recipient Port	Number of Tanks	STDEVs
Eastern Asia	Zhoushan	Vancouver(CAN)	854	0.235347
Eastern Asia	Rizhao	Vancouver(CAN)	615	0.255766
Eastern Asia	Qingdao	Vancouver(CAN)	608	0.251918
Eastern Asia	Shanghai	Vancouver(CAN)	395	0.253708
Eastern Asia	Lianyungang	Vancouver(CAN)	370	0.255404
Eastern Asia	Lanshan	Vancouver(CAN)	363	0.268970
Eastern Asia	Dangjin	Roberts Bank	275	0.237261
Eastern Asia	Caofeidian	Vancouver(CAN)	259	0.242195

3.9 in Appendix 3.4.

Overall, for the Canadian destination ports included in this study, it is clear that the lowest environmental distances (highest risk for NIS survival and establishment) are associated with source ports at similar latitudes in Europe, Eastern Asia and North America (Fig. 3.8).

3.4 Discussion

This study examined the temporal variation of temperature and salinity at ports worldwide and quantified the influence of this variation on environmental distance calculations as well as the corresponding risk for the introduction and establishment of aquatic NIS in Canadian waters. To do so, the use of monthly vs. annual average environmental data was considered within a baseline BWRA model. Except for some outliers in the Pacific region, the environmental distances based on monthly scale data generally become smaller in all regions (Fig. 3.6), demonstrating that the model using

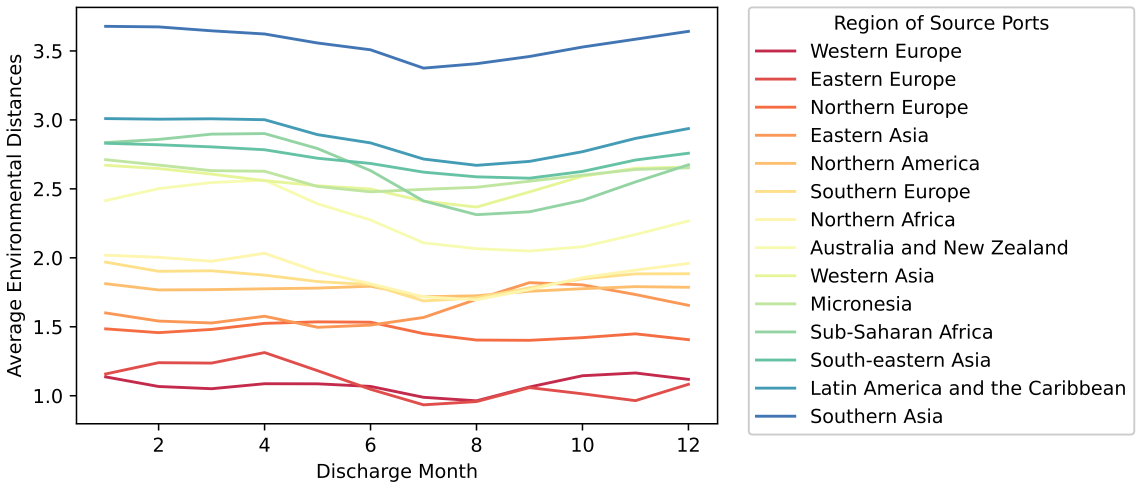


Figure 3.8: Average environmental distances across unique port-pair combinations during the 12 months of the year, grouped by source port region.

annual decadal average environmental data to inform environmental matching can underestimate risk of NIS survival and establishment in comparison to monthly data, at least for the combination of source-recipient port pairs occurring across Canada. Moreover, the distribution of the monthly vs. annual average environmental distances (Fig. 3.3) and the statistical comparison results (Table 3.2) follow the same pattern for both years, suggesting that the results are stable and can be generalized through time.

Spatial examination of differences between the monthly and annual average environmental distances shows that the cumulative risk across all tank discharges at individual ports can become much higher using monthly environmental data in the baseline BWRA model, with only a few ports experiencing a decrease in risk. Further, the assessment of monthly environmental distances for unique port pairs at fixed intervals throughout the year allows for an analysis of year-round risk variability for each port pair. Combined with the sources of ballast tanks, this study further explores the link between environmental conditions in the ballast water source regions and NIS survival and establishment risk.

Although the overall risk increases at most Canadian ports when using monthly environmental data, the regional statistics comparing monthly and annual average environmental distances show an uneven distribution of ballast tank discharges with higher and lower risk values. Some individual ports with increased and decreased risk are adjacent to each other because of the receipt of ballast water sourced from a specific location. For example, the only port with a markedly reduced risk, Campbell River, receives only a small number of tank discharges from two U.S. ports. In addition, the proportion of high-risk tanks discharged in a region can affect the result of statistical comparison (i.e., Wilcoxon effect size) for that region. For example, the ballast tank discharges in the Pacific contributed nearly 50% of all 'very high risk' extremes but represent only a small proportion of the total ballast tank discharges in the Pacific region, resulting in a small effect size in this region (Table 3.2).

In correspondence with a previous study which considered risk variation in discharge ecoregions [12], this study incorporates regional information for the source ports, enabling analysis of monthly risk variation in ballast water from specific sources and identification of additional risk patterns. The results indicate that for Canadian recipient ports, the overall invasion risk is higher when ballast water comes from ports at similar latitudes (e.g., Northern and Western Europe) and lower when coming from the tropical zone (e.g., Southern Asia and Latin America and the Caribbean) (Fig. 3.8). Combined with the temporal variation of environmental variables (Fig. 3.1), it can be also observed that the risk variation between port pairs often corresponds to larger interannual temperature variations - such as observed along the Mediterranean coast and northeast Asia. This finding is consistent with our hypothesis that ballast water from the temperate zone may have greater variability in assessed risk due to the large interannual variability in sea surface temperature in the temperate climate zone. At the same time, the monthly environmental distances also fluctuate markedly for port pairs without strong interannual temperature variation at the source port location, such as those in Australia and New Zealand, corresponding to higher risk when ballast discharges occur in Canada during the northern hemisphere's summer and autumn (Fig. 3.8). This pattern supports our hypothesis that the opposite seasons in the northern and southern hemispheres may create a higher risk for vessels departing in the southern hemisphere winter (northern hemisphere summer) to arrive at Canadian ports where the overall water temperature is cooler.

Port pairs with high variability in environmental distances and high shipping densities were examined, with the overlap being mostly from ports in Eastern Asia to ports on the west coast of Canada (Pacific region). The sizeable temporal variation in environmental distance between the two regions is possibly a result of: 1) large inter-annual variability in sea surface temperature in the northwest pacific [33] (i.e., the temperate climate zone of Eastern Asia); and/or 2) salinity fluctuations at the estuaries of large rivers [90] where the ports are densely distributed. Based on the seasonal variations observed in sea surface temperature (Fig. 3.1) and salinity (Fig. 3.2), risk changes are more likely to be influenced by temperature variations in temperate climate zones, as salinity has less seasonal variability along both the west coast of Canada and Eastern Asia.

Although there have been a number of previous studies implementing environmental matching in ballast water risk assessments [91, 92, 9, 10, 11], very few have analyzed the potential impacts of temporal variability in their models. Seebens et al. (2013) do demonstrate and discuss the occurrence of seasonal variability in the output of their global shipping invasion risk model, based on temporal variation in shipping intensity and temperature, though they do not quantify the difference and they continue to use annual average environmental data within their standard model. In the standardized assessment of monthly scale environmental risk conducted in this study, the factor of shipping intensity was excluded, leaving only the variability associated with the source and recipient ports environmental variables. However, in practical applications, considering shipping intensity is necessary since port pairs with moderate risk variation yet very high shipping intensity (i.e., high propagule pressure) deserve more attention than routes with significant risk variation and little shipping (i.e., low propagule pressure).

While this study examined the importance of temporal variation in environmental variables for BWRA, the results may extend more broadly to studies implementing species distribution models (SDM) to predict habitat ranges under current and future climate conditions based on environmental data associated with known occurrence/absence locations [93]. Both correlative and mechanistic SDM [94] have a strong reliance on environmental data, mainly climatic conditions. Many SDM have used environmental data at fixed spatial and temporal scales to define the distribution of species over spatial-temporal limits [95]. More specifically, annual data have been used to model the range of variation in environmental variables [96]. In response to changes in environmental variables, some studies have proposed a combination of climate change [97, 98, 99] and microclimate factors [100] to model species distributions. A recent study modeled the distribution of short-lived species using monthly historical data, and the proposed seasonal SDM can be better associated with habitability compared to conventional SDM [101]. Similarly, the results of our work suggests the use of finer-scale data reflecting the seasonal variability of environmental variables may achieve a more accurate prediction. Since some important variables, such as temperature, experience more seasonal variation on land than in the ocean, the use of monthly or quarterly data in SDM could have even greater influence on predictions of terrestrial species invasions and range shifts.

Several future research directions could be followed to tackle the remaining knowledge gaps and limitations of this study. Firstly, ballast water is known to be an important vector for introduction of NIS to freshwater ecosystems such as the Laurentian Great Lakes [102, 103]. Inland ports were excluded from this analysis due to the lack of environmental data near these ports in the World Ocean Atlas dataset; future work could include a seasonal assessment of environmental risk for ballast water discharges at inland ports if suitable data are available elsewhere. In addition, if finer scale global data are available for salinity, it would be desirable to further assess the temporal sensitivity of the environmental matching approach since salinity can fluctuate widely within a day at ports within estuaries subject to tidal influences. Moreover, the ballast tank records being fitted to models in this work span from 2019 to 2020, and are limited to discharges within Canadian waters. Although this research found similar patterns across two years, the generality of the patterns observed in this study could be examined across a wider geographic scope and time span. Nonetheless, the results of this study suggest future evaluations incorporating ballast water uptake and discharge dates (or ships' departure and arrival dates, if the former are not available) can provide a more sensitive assessment of risk reflecting seasonal variability compared to an annual average risk model.

Acknowledgement

We thank Nader Zare for assistance with data extraction from the Canadian Ballast Water Information System and for providing Python code for the baseline risk model. We sincerely appreciate the time and effort of three reviewers whose constructive comments and suggestions helped us to improve the quality of this manuscript.

Co-Authorship Statement

Amilcar Soares and Sarah A. Bailey conceived and supervised the project. Amilcar Soares and Ruixin Song designed the methodology. Ruixin Song collected and analysed the data, and led the writing of the manuscript. Yashar Tavakoli assisted with statistical analyses. Sarah A. Bailey and Amilcar Soares enhanced and facilitated the manuscript. All authors contributed to the study and approved the final manuscript.

Appendix: Hot Shipping Connections and Ballst Water Sources

Port pairs with more shipping connections

Table 3.4: Port pairs with large number of (more than 250) ballast tank connections in 2019-2020.

Source Port	Source Region	Recipient Port	Discharge Region	Number of Tanks
Boston(USA)	Northern America	Saint John(CAN)	Atlantic	2204
Portland(ME USA)	Northern America	Saint John(CAN)	Atlantic	1416
Providence	Northern America	Saint John(CAN)	Atlantic	897
Zhoushan	Eastern Asia	Vancouver(CAN)	Pacific	854
Stockton	Northern America	Vancouver(CAN)	Pacific	772
Redwood City	Northern America	Port McNeill	Pacific	660
Rizhao	Eastern Asia	Vancouver(CAN)	Pacific	615
Qingdao	Eastern Asia	Vancouver(CAN)	Pacific	608
Nantong	Eastern Asia	Vancouver(CAN)	Pacific	606
Los Angeles	Northern America	Vancouver(CAN)	Pacific	488
New York	Northern America	Come by Chance	Atlantic	408
Shanghai	Eastern Asia	Vancouver(CAN)	Pacific	395
Kashima	Eastern Asia	Vancouver(CAN)	Pacific	395
Chiba	Eastern Asia	Vancouver(CAN)	Pacific	393
New Haven	Northern America	Saint John(CAN)	Atlantic	393
Bayway	Northern America	Whiffen Head	Atlantic	387
Searsport	Northern America	Saint John(CAN)	Atlantic	380
Bayway	Northern America	Point Tupper	Atlantic	372
Lianyungang	Eastern Asia	Vancouver(CAN)	Pacific	370
Bayuquan	Eastern Asia	Vancouver(CAN)	Pacific	366
Lanshan	Eastern Asia	Vancouver(CAN)	Pacific	366
Baltimore, USA	Northern America	Halifax	Atlantic	313
Seattle	Northern America	Vancouver(CAN)	Pacific	306
Baltimore, USA	Northern America	Saint John(CAN)	Atlantic	304
Nagoya	Eastern Asia	Vancouver(CAN)	Pacific	302
Long Beach	Northern America	Port McNeill	Pacific	301
Providence	Northern America	Paspebiac	Atlantic	296
Long Beach	Northern America	Vancouver(CAN)	Pacific	288
Portland(OR USA)	Northern America	Vancouver(CAN)	Pacific	285
Portsmouth(NH USA)	Northern America	Saint John(CAN)	Atlantic	282
Mizushima	Eastern Asia	Vancouver(CAN)	Pacific	279
Dangjin	Eastern Asia	Roberts Bank	Pacific	275
San Francisco (USA)	Northern America	Vancouver(CAN)	Pacific	267
Caofeidian Port	Eastern Asia	Vancouver(CAN)	Pacific	259
Bucksport	Northern America	Saint John(CAN)	Atlantic	256

Distribution of the source ports

Fig. 3.9 shows the distribution of 602 source ports in this study. The different colors mark the 14 world regions in which the source ports are located, following the ISO-3166 standard [89].

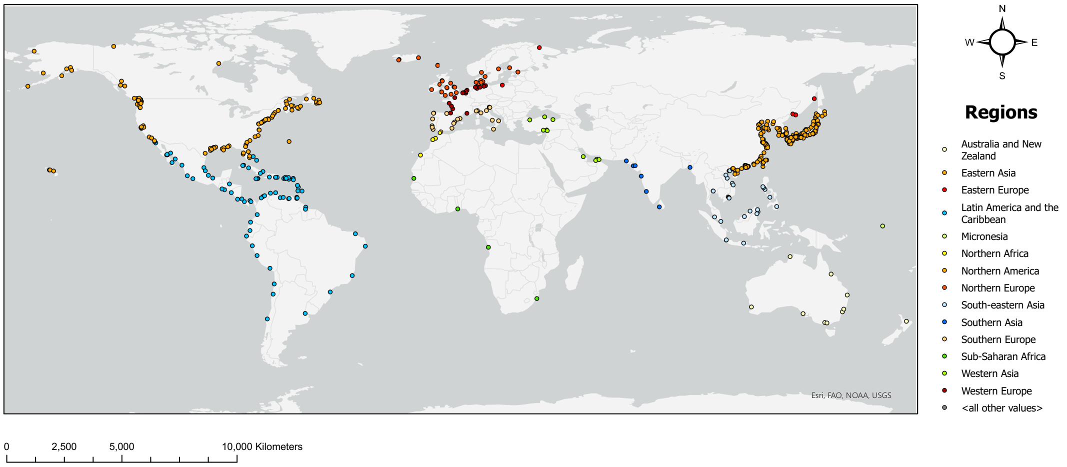


Figure 3.9: Regional distribution of 602 source ports categorized by 14 regions.

Chapter 4

Gravity-Informed Deep Learning Framework for Predicting Ship Traffic Flow and Invasion Risk of Non-Indigenous Species via Ballast Water Discharge

4.1 Introduction

Globalization has rapidly increased marine shipping activities in the last decades. According to a statistics report, container shipping has increased by 24 times in tonnage from 1980 to 2020 [104]. During this time, the environmental pollution caused by introducing Non-Indigenous Species (NIS) through shipping activities has been a subject of study regarding marine preservation. Ballast water, used to keep vessels in balance during travels, is listed as a major source of NIS pollution [29]. The introduction of NIS into different ecological regions due to ballast water discharge has been shown to cause significant damage to the local ecosystem, as it poses a severe threat to the biodiversity of affected areas [105, 77, 78]. In response to biological invasion issues, many studies about Ballast Water Risk Assessment (BWRA) have been conducted over the last decades to estimate the risk levels of carrying NIS in the ballast water tanks [106, 9, 10, 11, 13, 14, 46, 15, 34, 12, 107]. These works and tools rely on ship self-reports, such as those made available by the National Ballast Information Clearinghouse (NBIC) [108]. Ballast water reporting forms provide information on the water source and destination areas, allowing for an assessment of the environmental similarity between source and target locations of a vessel voyage, which is considered a significant indicator of invasion risk level [29]. However, the acquisition of ballast reports is limited at the global scale due to the various policies across different countries. Additionally, BWRA tools do not utilize alternative data sources that incorporate comprehensive shipping information.

Recent research has revealed a strong correlation between the introduction of non-indigenous species and the movement of ships through shipping networks. These studies [12, 40] have utilized data from the Automatic Identification System (AIS), a location tracking system on ships that allows them to share their positions in real-time [52]. This technology allows researchers to track individual/collective ships [109, 110, 111, 112], predict larger-scale shipping activities [113], and assess the risk of introducing NIS through ballast water [12]. AIS data has emerged as a promising source of information for studying bioinvasions in marine ecosystems. These studies analyze mobility flow by using Origin-Destination (OD) models that combine physics with statistical mechanics. The gravity model, inspired by Newton's *Law of Universal Gravitation*, measures the attractive force between two objects based on their masses and the distance between them [66]. The gravity-inspired OD models were introduced in early human mobility and migration studies [35, 36]. They rely on information about population size and distances between origins and destinations as features.

The gravity theory permeates many areas of study that go beyond mobility and migration, such as the spreading of epidemics [38, 39], commercial trading [37], communication [67] and shipping networks [40, 41] modeling. Although the gravity model has been widely used to model real-world problems, recent studies have shown that it may not be sufficient for capturing complex patterns in various scenarios [114]. Relying solely on mass and distance as the critical factors of the model could lead to failures in accurately representing patterns [115]. Nevertheless, the gravity model has been prevalent for many years and remains a popular tool for modeling various phenomena. Beyond gravity models, radiation absorption is another physics-inspired OD model to study mobility patterns [68, 116, 69]. Unlike gravity models, which draw inspiration from gravitational forces, radiation models are based on principles seen in radiation and absorption processes, also from physics. While gravity models contain adjustable variables that may be difficult to define, the radiation-absorption model simplifies this by emphasizing distance as the primary feature while considering the population density. A well-known application of the radiation model is to predict human movement toward locations influenced by employment indicators [68]. Further studies used field theory for abstracting vector fields of daily commuting flows [70], while others translated field theory-based mobility to deep learning models for achieving better interpretation of spatiotemporal features in mobility patterns [117].

The deep learning techniques discussed above in the mobility domain are a subset of the machine learning realm, and they have become increasingly popular in various applications due to their ability to recognize patterns by fine-tuning multiple parameters. These techniques have been used to forecast vessel trajectories in the ocean, predict patient trajectories in hospitals, track the spread of epidemics, and many other applications [118, 119, 120, 121]. Understandably, coupling machine or deep learning capabilities for pattern recognition with physics-inspired OD models can offer a higher capacity for capturing and predicting complex scenarios. For instance, a study in human mobility used deep learning methods with the gravity model [43]. The resulting composite model called the Deep Gravity Model, has expanded the standard feature set of conventional gravity models, which typically incorporated population size and distance between OD locations, to include a variety of parameters characterizing the origins and destinations such as land-use patterns and the presence of retail and healthcare amenities. A similar study proposed the use of machine learning models augmented with urban indicators to forecast the flow of commuters during worker migrations [42], which included data about labor, safety, and quality of life and work in the cities using censuses-collected data. The enriched feature set enabled both models to discern and capture the intricacies of human mobility across wider information spaces. Specifically, deep learning techniques have inherent learning mechanics characterized by feed-forward and backpropagation. These mechanics allow for adaptive weight assignment to individual variables. Therefore, this combination of deep learning and traditional OD models sets a promising premise for its application in maritime traffic mobility networks, as it allows intricate exploration of patterns at a granular level required for shipping network analysis.

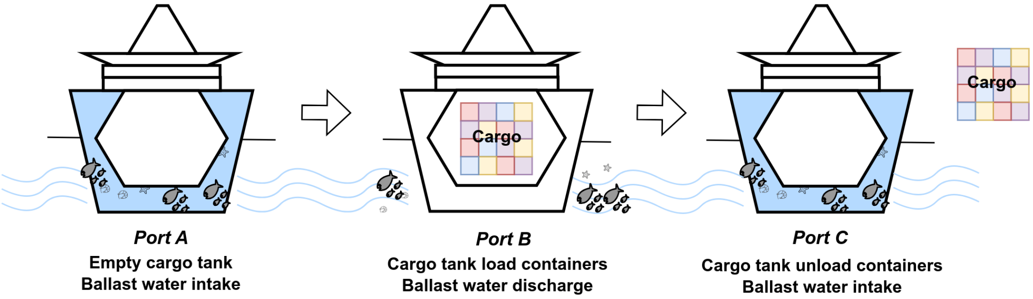


Figure 4.1: Non-indigenous species carried by ballast water during container shipping.

Along these lines, this paper focuses on predicting the OD fluxes of marine vessels to gain insight into global shipping behavior and their role in preventing NIS cases through BWRA. Ballast water is essential in maritime operations, particularly for container ships. As illustrated in Figure 4.1, while ballast water helps stabilize the ship's load, local species can enter the ballast tanks and travel long distances when the water is taken in and discharged. Therefore, forecasting shipping patterns is critical to understanding the risk of spreading NIS. To this end, we propose a gravity-informed model where the shipping fluxes are considered as "mass", and the vessel traffic flows are inversely proportional to the distances traveled by ships. We have enriched our model with relevant features from shipping activities, including bilateral trade data between countries [44] and graph metrics extracted from the global shipping network. Our deep learning model, known as the *Transformer Gravity*, relies on the transformer architecture [45] and is capable of capturing local and global data dependencies through self-attention mechanisms. This mechanism enables the model to weigh the importance of different parts of the input sequence, assigning varying degrees of attention depending on the relevance of each input part during generating the output. As a result, our model can discern and incorporate short- and long-term dependencies in vessel traffic flows, making it more sensitive to the complex and dynamic patterns in maritime vessel movements.

As part of our proposed framework, we have employed a machine learning classifier that proceeds the flow estimation process and filters out non-existing edges on a link-prediction binary classification task, a two-stage modeling process. This allows only highly probable flows, based on prior knowledge, to be fed into the gravity-based models, where the final flow estimations take place with the aid of the gravity-based model. In this regard, we have conducted experiments using the **(a)** Transformer Gravity, **(b)** Deep Gravity, and **(c)** shallower-layered variants of Deep Gravity Models. We utilized regression models in machine learning for performance comparison, and the same approach was used for the binary classification task. Our results demonstrate that the Transformer Gravity model significantly outperforms all the other approaches as it achieves an average Common Part of Commuters (CPC) [1] of 85.3%, representing an improvement of higher than 10% in the model output certainty in contrast to the Deep Gravity counterpart and over 20% improvement compared to other regression models. The results we have obtained are not only due to the proposed model but also

[1] See definition of CPC metric in Equation 4.4

to the proposed pipeline. We have significantly improved performance by incorporating prior knowledge about potential destinations and using the attention mechanism and the traditional gravity-informed model for mobility flow estimation.

Overall, we provide a consistent advancement in the gravity-informed flow estimation models that paved the way for ballast water risk assessment and management by enabling the forecasting of vessel mobility flows. That was possible mainly due to advancements in the model pipeline and architecture, which strongly rely upon key ocean and vessel mobility domain features. Specifically, data from global economic trades between countries and graph centrality metrics from port networks significantly contribute to achieving state-of-the-art results. This emphasizes the value of integrating shipping network analysis and trade information into vessel mobility flow predictions. Although gravity-informed models have limitations in capturing temporal dynamics [114], the annual patterns in ship traffic flows predicted by the Transformer Gravity model can provide references for shipping intensity in ballast water risk assessment. In addition, due to the versatility of mobility data, we believe that our framework and model can be utilized in various fields of research such as human mobility, urban transportation, and epidemic modeling.

4.2 Methods

4.2.1 Global Shipping Network

At the beginning of this study, we constructed a directed and weighted international shipping network based on global port visits data in the same way as in [54] from 2017 to 2019. Ports and shipping connections were represented as nodes and edges in the shipping network, and the World Port Index (WPI) [1] was also used for port identification.

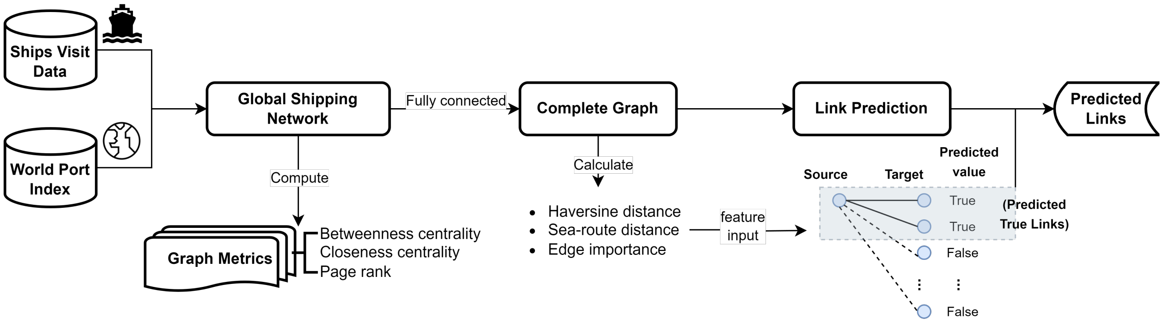


Figure 4.2: Pipeline for analyzing and predicting links in the global shipping network from 2017 to 2019.

The global shipping network can be described as $G = (V, E, W)$, where V represents the set of ports, E is the set of shipping routes connecting pairs of ports, and W is the collection of edge weights. In this context, each weight in W corresponds to the number of individual trips T_{ij} between port i and port j:

$$W = \{w_{ij} : w_{ij} = \sum_t T_{ij}(t) \, \forall (i, j) \in E\}$$

With the global shipping network defined, we performed network analysis to extract graph metrics as features for our proposed gravity-informed predictive model. We also conducted link prediction to identify potential origin-destination (OD) pairs within current shipping traffic, thus providing pre-knowledge to the predictive models. Figure 4.2 illustrates the pipeline of shipping network analysis and the link prediction process, which precedes the gravity-informed predictive models forecasting the ship traffic flows.

Graph Metrics Computation

We analyzed the shipping network by calculating various graph metrics for each node. These metrics included betweenness centrality, closeness centrality, and PageRank. The detailed equations for calculating these metrics are listed in Equations 4.14, 4.16, and 4.17 in the Methods. In this paper, betweenness centrality quantifies the frequency with which a port serves as an intermediary on the shortest paths between other ports — ports with high betweenness centrality play a critical role as bridges within the shipping network. Additionally, closeness centrality calculates the inverse of the sum of the shortest distances from a node to all other nodes — ports with a high degree of closeness are easily accessible from all other ports in the shipping network. Furthermore, PageRank [63] identifies essential nodes in a graph — ports with high metric values are more influential and likely to be frequently visited by ships from other important ports.

Link Prediction in the Shipping Network

Subsequently, during the shipping network analysis, we observed the presence of disconnected components. These disconnected segments pose a challenge for our proposed framework, which relies on integrating features across all possible destinations from each source port. Such disconnections can compromise the model's ability to generate accurate or well-defined predictions for shipping flows between isolated areas. Thus, we transformed the original network into a complete graph, denoted as G', and assigned a small value to weigh the new links. This action mitigates the issues arising from data sparsity and establishes a uniform data structure, thereby enhancing the robustness of the flow estimation framework. With the fully connected shipping network G', we proceed with the flow estimation framework by forecasting whether a trajectory exists by solving a link prediction problem in G', which can be framed as a binary classification task within machine learning models. This action gives concrete source-destination pairs well-prepared for building feature sets, modeling the gravity structure, and introducing the deep learning framework.

To perform link prediction, we first separated the shipping data from 2019 for testing and retained the data from 2017 and 2018 for training and validation. Then, we prepared features for the classification task. We calculated the Haversine distances (see Methods) between every pair of ports. However, Haversine distances only provide the geodesic approximation and cannot capture the real sea routes that the ships have traveled. Therefore, we also computed the sea-route distances between port pairs and obtained a more accurate representation using a Python package that models the shortest routes and calculates the sea route distances using historical trajectories [122]. Finally, we used these distances to evaluate the importance level I_{ij} for each edge $\langle i, j\rangle$ in the complete graph G'. Inspired by straightness centrality measuring the node connectivity by the straightness of the shortest distance [123], this metric combines the normalized flow size $\hat{w}_{ij}$ and the normalized Haversine distance $\hat{d_{ij}^E}$, using a small constant ϵ to prevent division by zero. Connections with more shipping flows and shorter distances are considered more important:

$$I_{ij} = \frac{\hat{w}_{ij}}{\hat{d_{ij}^E} + \epsilon},\ i, j \in V',\ i \neq j \tag{4.1}$$

Next, we incorporated the Haversine distance, sea route distance, and edge importance into the feature set to determine the existence of a trajectory between two ports. Upon labeling the real (true class) and pseudo (false class) links in the complete network G', we observed a significant class imbalance (2.3% real links and 97.7% pseudo links); consequently, we employed stratified sampling of the pseudo links based on their spatial distribution to balance the number of examples in each class in a manner that preserved the data's characteristics. Machine learning models were then engaged to perform binary classification, utilizing 75% of data from 2017 and 2018 for training and 25% for validation. During the training phase, a 5-fold grid search cross-validation was implemented to fine-tune the hyperparameters of each model. Finally, we tested the models on unseen data from 2019, reinforcing the validity of our approach.

These trajectories are then utilized as features of gravity-informed models for improving ship traffic flow prediction, such as seen in Figure 4.3, which allows for the environmental similarity analysis and the ballast water risk assessment discussed as follows.

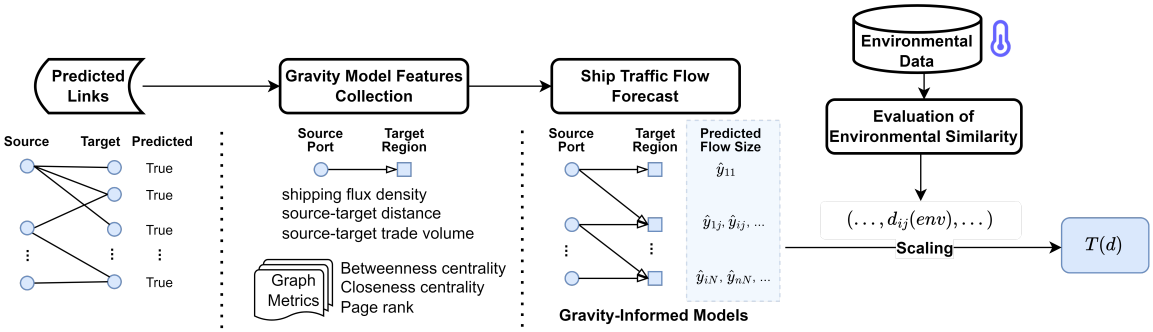


Figure 4.3: Experimental pipeline for predicting ship traffic flows with gravity-informed models and assessing environmental similarity for ballast water risk assessment (continuation from Figure 4.2).

4.2.2 Gravity Models

As in the laws of gravity, the gravity model describes the interaction between two entities proportionally to their masses and inversely proportional to their distance [66]. Over time, the model has been adapted to solve various practical problems, such as application in studying human mobility and migration patterns [35, 36], which follows the gravity basis that the number of commuters between two locations is related to the populations of the two locations and the physical distances. In addition, the gravity model also includes adjustable parameters that can be learned from historical data. Other popular applications include forecasting economic interactions [37], communication networks [67], and cargo shipping [40, 124, 41].

Even though the gravity model has prevailed for many years, its limitations can hardly be overlooked. First, the simple components of "masses" and "distances" cannot capture the connection with real mobility flows and lack comprehensive factors that can represent specific scenarios [68], such as the city infrastructure in urban mobility studies and the global trade pattern impacting cargo shipping flows. Also, the lack of limits on the flow increase can lead to the predicted flow size larger than the source "masses", making the model predictions unreasonable and challenging to interpret [68, 42]. Additionally, it is difficult to deal with multiple adjustable parameters in the model for the prediction without enough previously observed data.

Alternative solutions have been proposed to overcome the limitations of gravity models in predicting mobility flows. In this sense, the radiation model offers a parameter-free approach which resolves the problem of having multiple parameters in the gravity model. Besides, it limits the number of flows by introducing the total number of individuals departing from the source location [68, 69]. Moreover, different studies used multiple relative factors, such as employment and urban infrastructure, that capture more important characteristics to improve prediction accuracy and adaptability of the model [42, 43]. The evolving nature of artificial intelligence and the rise of large language models have brought about new technologies. Therefore, improving the state of the art by using more capable technologies over new arrangements and combinations of data is essential. This becomes more evident when considering evolving factors such as climate change and recurrent anthropogenic effects observed on the oceans.

Deep Gravity

Deep Gravity is a recent study in urban mobility that integrates the gravity model with deep neural networks [43]. Unlike traditional gravity models that rely on maximum likelihood estimation for parameter tuning, Deep Gravity utilizes a cross-entropy loss function as a substitute for maximum likelihood estimation. This innovative approach effectively adapts the gravity model to a neural network framework. Also, it departs from the traditional multi-parameter structure of gravity models; instead, it incorporates the populations, distances, and infrastructures into the feature set and integrates data from various sources.

However, the simplicity of the multilayer perceptron (MLP) structure in Deep Gravity presents certain limitations. Specifically, the complex multivariate relationships inherent in mobility flows can be challenging to capture accurately through the composition of multiple linear functions. Moreover, the deep architecture composed of fifteen linear layers stacked in sequence demands many model parameters. Thus, it requires more computational resources and training time to provide a viable solution. While Deep Gravity utilizes a feed-forward fully-connected multilayer architecture, the underlying gravity model can be simplified to contain one or a couple of layers. In this study, we employed multi-layered models configured with 3, 9, 12, and 15 layers as benchmarks for comparison against our proposed framework that leverages Transformers on Deep Gravity Models.

4.2.3 Transformer Gravity

In this study, we incorporate the self-attention mechanism of the Transformer architecture [45] into our proposed framework. Compared to the conventional MLPs structure, the self-attention mechanism can inspect the input sequence and weigh to identify and prioritize the most relevant elements for generating the output, and this characteristic enables our model to capture crucial dependencies in vessel mobility flows effectively. Additionally, the self-attention mechanism accomplishes high performance with fewer parameters, making it a computationally efficient model. This section shows how we model the *Transformer Gravity*, combining the characteristics of the Gravity Model and the self-attention mechanism.

Problem Definition

Based on the pairs of source-destination ports obtained from the previous link prediction step, we encode the destination ports into 17 geographical regions according to the ISO-3166 standard [89], where ports within regions are expected to have a similar set of organisms and, therefore, share similar habitat. Over the encoded representation of regions, we now aim to estimate the sizes of shipping mobility flows between each source-destination pair (P_i, P_j), where P_i is the source port and P_j is the destination port that pertains to a unique geographical region. A ship departing from a source port may have one or more destination ports in the same or different regions, and following the Deep Gravity [43] method, the goal is to estimate probabilities of the ships traveling to these geographical regions, becoming a multiclass classification task.

Predict target: We compile a set of 10 features from various sources for each pair (P_i, P_j), such as shipping fluxes at ports, geodesic distances between source and destination, global economic trade volume, the graph metrics computed with the global shipping network, and others; detailed information about these extracted features is provided in Table 4.3 in the Methods. We represent the feature vector for each source-destination pair as $x_{ij} = \langle m_1, m_2, \dots, m_{10} \rangle$. Given the ships from each source port travel to multiple destination regions, these feature vectors are aggregated into a single data sample $X_i = \{x_{i1}, x_{i2}, \dots, x_{iN}\}$, where N is the number of destination regions in the data sample, and $1 \leq N \leq 17$. Each destination region is represented as a class, so the prediction has N classes for a sample. Since samples of varying lengths cannot be wrapped to a tensor for batch processing, we set the batch size to 1 for model input. Using $\hat{y}_{ij}$ as the estimated size of the mobility flow between (P_i, P_j), which is the target of the prediction, we have:

$$\hat{y}_{ij} = O_i \cdot p_{ij} \equiv O_i \frac{e^{f(x_{ij})}}{\sum_{k=1}^{N} e^{f(x_{ik})}} \tag{4.2}$$

where O_i represents the total number of ships departing from source port i. p_{ij} is the probability of ships traveling from source port i to the destination region j, and $f(x_{ij})$ is the model output of feature vector x_{ij}.

Loss function: We used the cross-entropy loss function for the model optimization process, defined as:

$$L\left(\hat{y}_{ij}, y_{ij}\right) = -\sum_{i=1}^{M} \sum_{j=1}^{N} y_{ij} \cdot \ln\left(\frac{e^{f(x_{ij})}}{\sum_{k=1}^{N} e^{f(x_{ik})}}\right) = -\sum_{i=1}^{M} \sum_{j=1}^{N} y_{ij} \cdot \ln\left(\frac{\hat{y}_{ij}}{O_i}\right) \tag{4.3}$$

The function presents the total loss between the predicted flows $\hat{y}_{ij}$, and the real flows

y_{ij} for all the N destination regions from M source ports. The *log-softmax* function is applied to the model output $f(x_{ij})$, and the loss function in terms of p_{ij} is obtained by replacing the log-term by Equation 4.2 divided by O_i.

Evaluation metric: The *Common Part of Commuters* (CPC) [65, 125, 126] is designed to measure the similarity between two sets of data, which could represent various aspects such as the volume of commuters, traffic, or trade between different locations. The metric calculates how much overlap there is between the predicted values $\hat{y}_{ij}$ and the actual values y_{ij}. In the context of Equation 4.4, M represents the number of source ports and N the number of destination regions. The values $\hat{y}_{ij}$ and y_{ij} correspond to the flow of vessels from source port i to destination region j, based on a model's prediction and the actual observed values, respectively. A high value of the CPC, in this case, means that there is a large overlap between the predicted and actual datasets. Specifically, it would indicate that the predictions accurately capture the true data distribution patterns, with most predictive quantities closely matching the actual quantities. Contrarily, a low value of the CPC would suggest that there is little overlap between the predictions and the actual data, indicating that the model's predictions diverge significantly from the observed data, which could be due to underprediction or overprediction in various parts or a general misalignment of the model with the reality. Accordingly, CPC considers the minimum common value between the predicted and actual data for each pair of source and destination ports, measuring the intersection over the values union:

$$CPC(\hat{y}_{ij}, y_{ij}) = \sum_{i=1}^{M} \frac{2\sum_{j=1}^{N} \min(\hat{y}_{ij}, y_{ij})}{\sum_{j=1}^{N} \hat{y}_{ij} + \sum_{j=1}^{N} y_{ij}} \tag{4.4}$$

For a better evaluation process, besides the CPC, we included the Normalized Root Mean Square Error ($NRMSE$) — lower is better — and Pearson Correlation Coefficients ($Corr.$) — higher is better — to measure the normalized errors and the correlation between the predictions and observations, defined as:

$$NRMSE(\hat{y}_{ij}, y_{ij}) = \sum_{i=1}^{M} \frac{\sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_{ij} - \hat{y}_{ij})^2}}{\max(y_{ij}) - \min(y_{ij})} \tag{4.5}$$

$$Corr.(\hat{y}_{ij}, y_{ij}) = \sum_{i=1}^{M} \frac{\sum_{j=1}^{N}(y_{ij} - \overline{y}_{ij})(\hat{y}_{ij} - \overline{\hat{y}}_{ij})}{\sqrt{\sum_{j=1}^{N}(y_{ij} - \overline{y}_{ij})^2 \sum_{j=1}^{N}(\hat{y}_{ij} - \overline{\hat{y}}_{ij})^2}} \tag{4.6}$$

Model Framework

Our proposed Transformer Gravity model is composed of three main components: (1) the input embedding layer, which maps the input feature vectors to a higher-dimensional space that is compatible with the Transformer architecture; (2) the multilayer Transformer encoder, which involves the self-attention and feed-forward blocks that process the embeddings to capture complex relationships between input features; and, (3) the output linear layer, which maps the processed embeddings to the target flow predictions, computes loss and CPC and performs backpropagation based on the loss values. Figure 4.4 presents the model pipeline using two stacked Transformers modules and provides a glance at the layer's relationships.

Linear Embedding. The embedding layer takes the input sample, which is a sequence of feature vectors represented as $\{x_{i1}, x_{i2}, \ldots, x_{iN}\}$. It then maps each vector into a higher-dimensional space using a linear transformation that involves a weights matrix and a bias vector. The result of this transformation is the feature embedding $\mathbf{z}_0$ for each vector x_{ij}, which can be obtained following the subsequent calculation:

$$\mathbf{z}_0 = x_{ij} \cdot {W_0}^\top + b_0,\ x_{ij} \in \mathbb{R}^{1\times n},\ W_0 \in \mathbb{R}^{d\times n},\ b_0 \in \mathbb{R}^{1\times d} \tag{4.7}$$

The input vector x_{ij} with 10 features is represented by W_0 (the weight matrix) and b_0 (the bias vector). This input vector is then embedded into a 64-dimensional space,

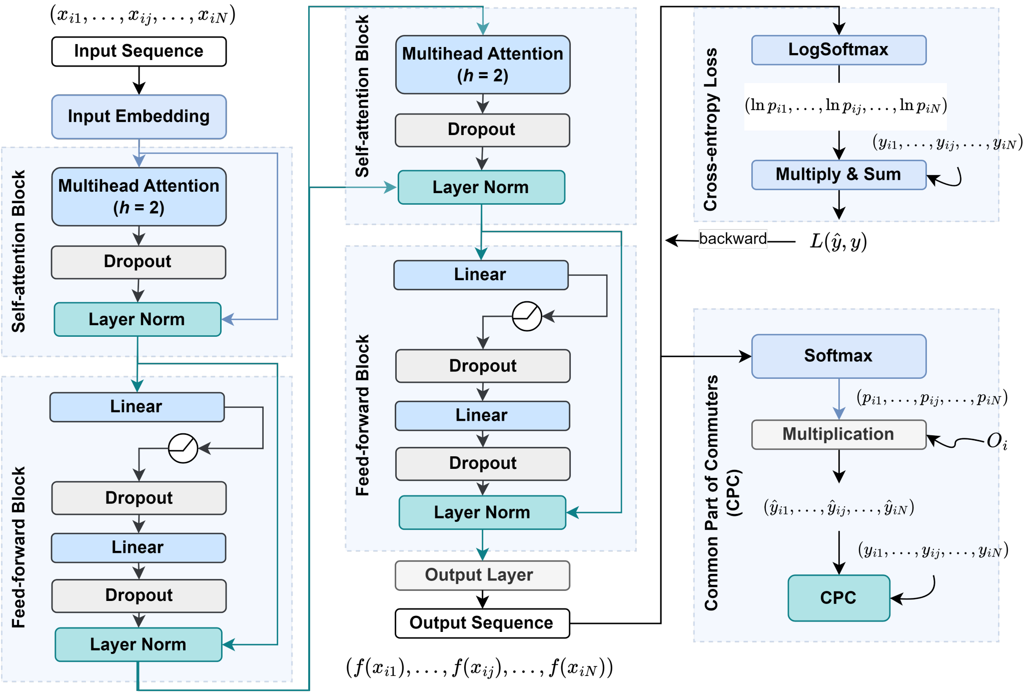

$(x_{i1}, \dots, x_{ij}, \dots, x_{iN})$
Input Sequence
Input Embedding
Self-attention Block
Multihead Attention (*h* = 2)
Dropout
Layer Norm
Feed-forward Block
Linear
Dropout
Linear
Dropout
Layer Norm
Self-attention Block
Multihead Attention (*h* = 2)
Dropout
Layer Norm
Feed-forward Block
Linear
Dropout
Linear
Dropout
Layer Norm
Output Layer
Output Sequence
$(f(x_{i1}), \dots, f(x_{ij}), \dots, f(x_{iN}))$
Cross-entropy Loss
LogSoftmax
$(\ln p_{i1}, \dots, \ln p_{ij}, \dots, \ln p_{iN})$
$(y_{i1}, \dots, y_{ij}, \dots, y_{iN})$
Multiply & Sum
backward
$L(\hat{y}, y)$
Common Part of Commuters (CPC)
Softmax
$(p_{i1}, \dots, p_{ij}, \dots, p_{iN})$
Multiplication
O_i
$(\hat{y}_{i1}, \dots, \hat{y}_{ij}, \dots, \hat{y}_{iN})$
$(y_{i1}, \dots, y_{ij}, \dots, y_{iN})$
CPC

Figure 4.4: Framework of the Transformer Gravity model.

resulting in an embedded output $\mathbf{z}_0 \in \mathbb{R}^{1\times d}$. Subsequently, the embedded output is passed to the multi-head attention encoder layers as the input.

Multi-Head Attention. A multi-head attention encoder comprises a multi-head self-attention mechanism and a feed-forward network, followed by layer normalization (as illustrated in Figure 4.4). Within each self-attention head, the input $\mathbf{z}_0$ is transformed into queries Q_h, keys K_h and values V_h using the weight matrices W_Q, W_K and W_V, respectively. Self-attention then calculates $Head_h = softmax\left(\frac{Q_h \cdot {K_h}^\top}{\sqrt{d_k}}\right) \cdot V_h$, where $d_k = \frac{d}{h}$ is the dimension of the queries Q_h and keys K_h and is used to scale the product $Q_h \cdot {K_h}^\top$. Multi-head attention combines all heads and linearly transforms the concatenation to produce $\mathbf{z}_1 \in \mathbb{R}^{1\times d}$:

$$\mathbf{z}_1 = Concat(Head_1, \dots, Head_h) \cdot W_C, \quad W_C \in \mathbb{R}^{hd_v \times d} \tag{4.8}$$

In our experiment, we define the number of heads as $h = 2$, and the heads run operations in parallel.

Layer Normalization. After the multi-head attention layer, there is a dropout layer that randomly sets a certain percentage of elements to 0. The dropout ratio p is set to 0.1 as per Equation 4.9. Next, a skip connection is applied to add the input features $\mathbf{z}_0$ to the output of the dropout layer $\mathbf{z}_{dropout}$ before the self-attention block. This helps to retain the information from the input features and prevent vanishing gradients during backpropagation. The output of this connection, $\mathbf{z}_{skip}$, is then normalized using Equation 4.9, where μ is the mean and σ is the standard deviation of $\mathbf{z}_{skip}$ with a small bias. The affine parameters α and β are initialized as 1 and 0, respectively, and can be optimized during the training process.

$$\begin{aligned} \mathbf{z}_{skip} &= \mathbf{z}_0 + \mathbf{z}_{dropout} \equiv \mathbf{z}_0 + Dropout(\mathbf{z}_1, p) \\ \mathbf{z}_2 &= LayerNorm(\mathbf{z}_{skip}) \equiv \frac{\mathbf{z}_{skip} - \mu}{\sigma} \times \alpha + \beta \end{aligned} \tag{4.9}$$

Feed-Forward Network. After the multi-head attention block processes the input, the resulting output is fed into a feed-forward neural network composed of an MLP structure. The connectivity of each layer in the feed-forward block's structure is illustrated in Figure 4.4. We formulate the output vectors from the layers using corresponding weight updates in Equation 4.10. Similar to $\mathbf{z}_{skip}$, a skip connection adds the vector $\mathbf{z}_2$ to the output $\mathbf{z}_4$ to preserve information from the self-attention block.

$$\begin{aligned} \mathbf{z}_3 &= Dropout\left(ReLU(\mathbf{z}_2 \cdot W_1^\top + b_1), p\right) \\ \mathbf{z}_4 &= Dropout\left((\mathbf{z}_3 \cdot W_2^\top + b_2), p\right) \\ \mathbf{z}_5 &= LayerNorm(\mathbf{z}_4 + \mathbf{z}_2) \end{aligned} \tag{4.10}$$

Training and Optimization. Our Transformer Gravity model has three transformer encoder layers stacked together to capture complex input embedding dependencies, but the number of stacked layers can be changed to match different requirements and needs. The output value $\mathbf{z}_5$ is obtained by passing the output of Equation 4.10

through a second and later third multi-head attention and feed-forward network block. The output value of the model is denoted as $f(x_{ij})$, and it produces a sequence of output values for a single data sample with a length of N. This sequence is then applied to a *softmax* function to produce probabilities $\{p_{i1}, p_{i2}, \ldots, p_{ij}, \ldots, p_{iN}\}$ for N classes. The predicted flow sizes $\{\hat{y}_{i1}, \hat{y}_{i2}, \ldots, \hat{y}_{ij}, \ldots, \hat{y}_{iN}\}$ for each destination are obtained by multiplying these probabilities with the total outflows O_i from the source port, as given in Equation 4.2. The loss for every sample is computed using Equation 4.3, and these losses are collected to derive the total loss. The model's parameters are updated with each loss by processing a single sample. The summed CPC across all samples is calculated using Equation 4.4. After each training epoch, the summed CPC is divided by the number of samples M (*i.e.*, the number of source ports) to obtain the average CPC of that epoch. During training, we used the Adam optimizer with L_2 regularization on the weights and reduced the learning rate by a factor of 0.1 when there was no improvement with the validation CPC after 10 epochs. We applied early stopping when there was no improvement with the validation CPC after 20 epochs to prevent overfitting and improve training efficiency. In addition, we used 5-fold cross-validation during training to ensure the reliability of the results.

4.2.4 Risk Assessment of NIS through Ballast water

As part of our study, we conducted a final experiment using the models we proposed on the global shipping network. Our goal was to assess the risk of NIS invasion associated with shipping flows using the BWRA decision tool, also employed by Transport Canada [34]. We aimed to demonstrate how our proposed model can improve real-world risk analysis, providing better information for regulating the oceans and making policies. The tool takes into consideration environmental conditions such as sea surface temperature and salinity at the locations where ballast water is taken and discharged. We gathered environmental variables, including minimum, maximum, and annual temperature, and annual salinity at these locations, and compiled them into a vector called $\nu_i = \langle t_{(i)_{min}}, t_{(i)_{max}}, t_{(i)}, s_{(i)} \rangle$. Through the vector ν_i, we proceed to calculate the environmental distance using the element-wise Euclidean distance calculation:

$$d_{ij}(env) = \sqrt{\sum_{k=0}^{|\nu|} (\nu_{i_k} - \nu_{j_k})^2} \tag{4.11}$$

Based on Equation 4.11, a smaller $d_{ij}(env)$ value suggests a higher environmental similarity between the origin and destination, typically meaning a higher risk of NIS invasion via ballast water and vice versa.

Following the approach described above, we first integrated port data from *World Port Index* [1] with the environmental conditions in *Global Port Environmental Data* [127] to match the environmental variables with the ports. We then assigned these variables to each port involved in the ship movement of 2019, thus forming the environmental vector pairs (ν_i, ν_j) for each pair of sources and destinations. Using Equation 4.11, we calculated the environmental distances for these pairs. In order to account for the differences in environmental distances across the shipping network, we included the frequency of each computed distance. This frequency is directly proportional to the predicted size of the shipping flow, which helps to adjust the environmental distance data to match the volume of shipping traffic. As illustrated in the pipeline diagram (Figure 4.3), we scaled the environmental distances using shipping volume from the Transformer Gravity model (denoted as $T(d)_{TG}$), the Deep Gravity model ($T(d)_{DG}$) and compared these with the scaled environmental distances based on actual shipping flows from 2019 (represented as $T(d)_{true}$). This approach was used to evaluate the dissimilarity of risk assessment results associated with the predicted and real shipping flows.

4.3 Results

4.3.1 Global Shipping Network Analysis

Figure 4.5 provides an overview of global shipping connections in the three years from 2017 to 2019 and the distribution of ports participating in these shipping activities.

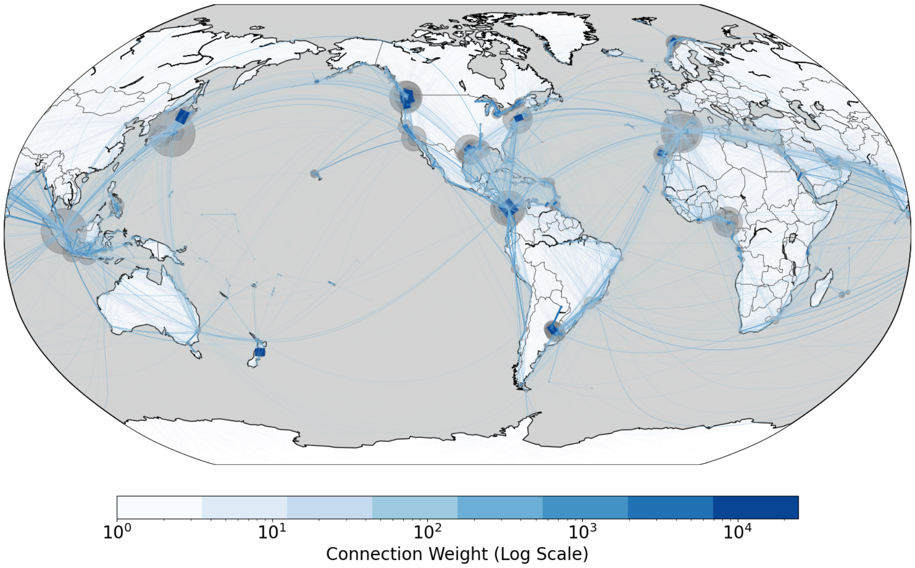


Figure 4.5: Global shipping network joining data from 2017 to 2019.

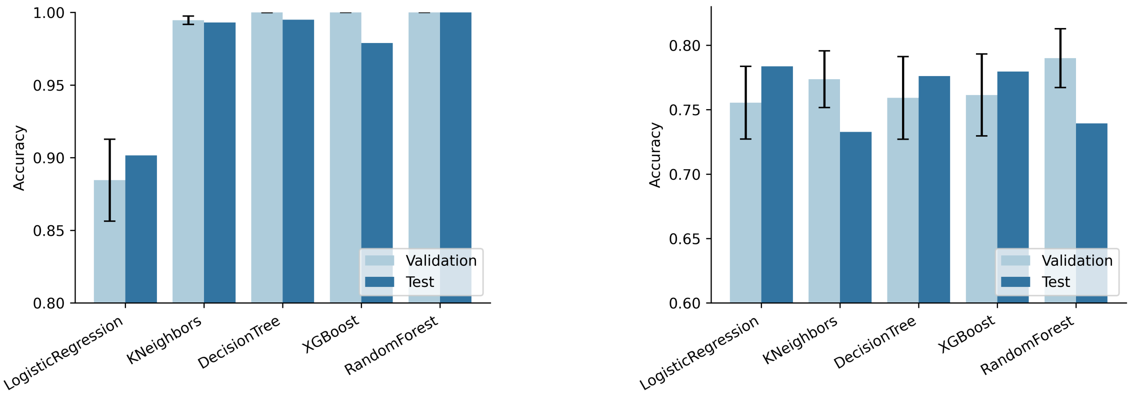
After calculating three different graph metrics, we noticed that they all showed positive non-linear correlations with each other. This means that as one metric increased, the others also tended to increase. Although each metric has a distinct interpretation, we found five ports from different bodies of water that had high values in all three metrics (see Table 4.1). These ports are particularly noteworthy because they excel in multiple areas. Many of the most influential ports are located at crucial marine traffic junctions that connect different oceans, such as the Gulf of Suez, the Gulf of Panama, and the Strait of Malacca.

Table 4.1: Ports with the highest graph metric values in betweenness centrality ($C_B(i)$), closeness centrality ($C_C(i)$), and page rank (P_{ij}) with port information [1], including water body, country, and region.

Port Name	$C_B(i)$	$C_C(i)$	P_{ij}	Water Body	Country	Region
Keppel	0.021	0.562	0.012	Strait of Malacca	Singapore	South-eastern Asia
Europa Point	0.018	0.553	0.012	Strait of Gibraltar	Gibraltar	Southern Europe
Puerto Cristobal	0.018	0.539	0.007	Caribbean Sea	Panama	Latin America & Caribbean
As Suways	0.018	0.532	0.003	Gulf of Suez	Egypt	Northern Africa
Balboa	0.018	0.523	0.006	Gulf of Panama	Panama	Latin America & Caribbean

4.3.2 Link Prediction in the Shipping Network

Figure 4.6a displays the 5-fold cross-validation and test accuracy for various classification models. It can be observed that apart from Logistic Regression at 89%, all other models achieved high accuracy (above 98%) on both the validation and the test datasets. We further investigated the discrepancies between the model with the highest performance (Random Forest at 100%) and the model with the lowest accuracy (Logistic Regression at 89%). The result revealed that the links primarily missed by Logistic Regression were those with only 1 or 2 trips and, to a lesser extent, 3-trip links. We also observed that the high accuracy of some models could be attributed to the inclusion of normalized edge weights w_{ij} in G' (*i.e.*, the normalized number of trips) in the edge importance calculation as per Equation 4.1. Consequently, we ran predictions without edge importance as a feature to mitigate this effect for comparison. Figure 4.6b presents the validation and test accuracy when relying solely on Haversine and sea route distance as features. The accuracy of these models ranges from 75% to 79%. Notably, the models' performance decreases by 16% to 20% compared to models that use edge importance as a feature. Retaining critical details from trade fluxes directly proportional to the link's existence is the key function of the edge importance feature. However, we have observed that more than the edge importance is needed to create a self-contained inference model. The challenge lies in selecting a model that can utilize the extra information without overfitting the data.



(a) (b)

Figure 4.6: 5-fold cross-validation average and test accuracy of classifiers in the trajectory link prediction task. (a) Performance of the classification task incorporates Haversine distance, sea route distance, and edge importance as features. (b) Performance of classification task without the edge importance features.

After analyzing the results from the model shown in Figure 4.6a, we observed that the Random Forest model has a higher accuracy. However, the Logistic Regression model acts as a filter by removing some 1- and 2-trip trajectories before predicting ship traffic flow. These trajectories are considered unstable and unreliable for predicting future mobility flow and may be outliers. Therefore, the Logistic Regression model is used to identify and remove these trajectories, leaving behind only those predicted to have more stable traffic flow between origin-destination pairs.

4.3.3 Shipping Flow Prediction

During the experimentation, we used the proposed Transformer Gravity model with 1, 3, and 5 Transformer layers. This study compared the original 15-layer Deep Gravity model with multi-layered variants of 3, 9, and 12 layers. Additionally, we experimented with machine learning regression models, such as linear-based, tree-based, and boosting-based models, as detailed in Table 4.2. The data from 2017 and 2018 formed the training data, while the 2019 data was used for testing.

Table 4.2 shows the mean, maximum, and minimum values of the models' CPC across the five validation folds. Our proposed Transformer Gravity, particularly the ones with 3 and 5-layer configurations, have achieved the best performance in cross-validation with a mean CPC of 0.864, which marks a 10.5% improvement over the top-performing 3-layer Deep Gravity variant ($CPC\ of$ 0.782), and a 13.2% improvement compared to original Deep Gravity model ($CPC\ of$ 0.763) and over 49.7% to other machine learning models, whose mean CPCs ranged from 0.474 to 0.577. Meanwhile, we noticed that compared with the original Deep Gravity model, its shallower-layered variants show a better performance, evidenced by metrics in both validation and test. However, none of these Deep Gravity variants can exceed the lowest mean CPC performance ($CPC\ of$ 846 using a single layer) of the Transformer Gravity model. The gap between Transformer Gravity's performance and other models indicates that our model's predicted shipping flows have a larger resemblance to the real shipping flows, highlighting our model's ability to reflect the real shipping mobility flows. This high similarity is also reflected in the other two metrics, $NRMSE$ and $Corr.$ The 3-layer Transformer Gravity model achieved the lowest error rate ($NRMSE\ of$ 0.080) and the highest correlation ($Corr.\ of$ 0.977) with actual shipping flows. These results suggest the Transformer Gravity model's superior performance in predicting global ship traffic flows over the Deep Gravity model.

After analyzing CPC values in Table 4.2, it can be observed that the Transformer Gravity models show a more stable performance than the competing models. The difference between the highest and lowest CPC values for the Transformer Gravity models ranges from 0.015 to 0.027, which is lower than other gravity-based models (0.019 $\sim$ 0.049) and machine learning models (0.164 $\sim$ 0.233). This indicates that the variance in CPC is lower when the flow prediction is done with the Transformer

Table 4.2: Performance evaluation of the Transformer Gravity, Deep Gravity and its shallower-layered variants, and other baseline models. The cross-validation results present the mean (CPC_{mean}), maximum (CPC_{max}), minimum (CPC_{min}), 5-fold CPC standard deviation (STD_{cpc}), mean Normalized Root Mean Square Error ($NRMSE$), and the mean Pearson Correlation Coefficients ($Corr.$) across five folds. Test results show the performance of the trained models on the entire test dataset. Details on the number of layers and parameters are provided for each deep learning model.

Model Name	*Layers*	*5-Fold Cross-Validation*						*Testing*			*Parameters*
		CPC_{mean}	CPC_{max}	CPC_{min}	STD_{cpc}	$NRMSE$	$Corr.$	CPC	$NRMSE$	$Corr.$	
Linear Regression	—	0.474	0.564	0.353	0.072	0.327	0.457	0.570	0.327	0.563	—
Decision Tree	—	0.518	0.625	0.438	0.092	0.404	0.370	0.644	0.398	0.517	—
Random Forest	—	0.577	0.685	0.504	0.071	0.309	0.545	0.675	0.285	0.692	—
Extra Tree	—	0.573	0.686	0.483	0.077	0.311	0.535	0.699	0.255	0.767	—
Gradient Boosting	—	0.557	0.654	0.490	0.065	0.306	0.555	0.654	0.295	0.665	—
XGBoost	—	0.552	0.696	0.463	0.088	0.320	0.521	0.664	0.282	0.702	—
LightGBM	—	0.574	0.706	0.482	0.083	0.307	0.554	0.684	0.279	0.707	—
CatBoost	—	0.559	0.674	0.472	0.075	0.311	0.544	0.654	0.288	0.683	—
Deep Gravity [43]	3	**0.782**	**0.797**	**0.766**	**0.006**	**0.209**	**0.833**	0.787	0.238	0.802	52,353
Deep Gravity [43]	9	0.769	0.776	0.751	0.008	0.220	0.812	0.783	**0.243**	0.797	249,985
Deep Gravity [43]	12	0.767	0.775	0.756	0.006	0.219	0.814	**0.790**	0.239	**0.803**	348,801
Deep Gravity [43]	15	0.763	0.778	0.729	0.016	0.222	0.808	0.752	0.273	0.747	447,617
Transformer Gravity	1	0.846	0.856	0.829	0.002	0.161	0.898	0.834	0.200	0.865	51,212
Transformer Gravity	3	**0.864**	0.870	0.852	**0.002**	**0.080**	**0.977**	**0.848**	**0.187**	**0.882**	101,644
Transformer Gravity	5	**0.864**	**0.87**1	**0.856**	0.008	0.107	0.953	0.836	0.208	0.858	152,076

Gravity model. The variance can be attributed to the distribution of data samples across folds, as a random seed is set for the sample assignment. This way, the Transformer Gravity model's performance is more consistent across different folds, meaning that given historical shipping data within the same temporal period, the Transformer Gravity model is more likely to predict shipping traffic flows with stable performance, regardless of the shipping traffic from which subsets of the data are sampled from.

On the right side of Table 4.2, we show the test results obtained by models trained on all the data from 2019 at once. The best CPC was produced by the 3-layer Transformer Gravity model ($CPC\ of$ 0.848), which indicates that the overlap ratio between predicted and actual shipping traffic flows was very high for an unseen scenario that spanned a year. This benefits us, as it helps us obtain more accurate shipping patterns, which we can use to evaluate the shipping intensity for ballast water risks in the future. Moreover, our Transformer Gravity model had a Pearson Correlation Coefficient of 0.882. This value indicates a strong linear relationship between predicted and actual flows, demonstrating the model's performance.

Table 4.2 shows that several baseline models have higher test CPCs than their best cross-validation CPCs. This difference in performance can be attributed to the distinct feature distributions in the test data compared to the cross-validation fold data. Additionally, the spatial and temporal dependencies intrinsic to the shipping data can also affect these evaluation results. Our training set includes shipping data from 2017 and 2018, while the test set comprises the entire year of 2019. Since both data sets represent complete years, the test data is expected to be more closely aligned with the training data in spatio-temporal distribution. This similarity can potentially explain the higher test performance than the validation results observed for these models. However, the evaluation of the Transformer Gravity model shows a better result in validation sets, which follow a more conventional pattern and have minimal variance over the cross-validation folds. This suggests that our model has learned more intrinsic connections from data features and is robust enough to overcome the impact potentially caused by the different distributions of datasets, demonstrating the model's generalizability and reliability in predicting mobility flow patterns.

Table 4.2 also reveals that the performance of the Transformer Gravity model varies with the addition of layers. An increase in the number of the Transformer encoder layers initially enhances the model's performance, as evidenced by the results of the 3-layer Transformer Gravity model compared with the single-layer model shown in the table. However, this performance increase is not unlimited. Comparing the 3-layer Transformer Gravity with the 5-layer one, as the model grows deeper with more integrated parameters, its performance tends to reach a plateau. Therefore, while stacking Transformers can be beneficial, it is important to optimize the number of layers to maximize the model's performance.

Similarly, the Deep Gravity model also needs an optimized configuration. The original 15-layer Deep Gravity model includes 256 hidden dimensions for layers 1 to 5 and 128 dimensions for layers 6 to 15. Due to the relatively large performance variation among its validation folds (CPC_{max} of 0.778 and CPC_{min} of 0.729), we adjusted the number of layers and experimented with shallower versions of the model for comparison. We followed the 1:2 ratio of the 256 and 128 dimensions in the original Deep Gravity to configure the 3, 9, and 12-layer variants. As per Table 4.2, we observed that Deep Gravity models with fewer layers delivered better performance than the original deep-layered model consisting of 15 layers. However, the performance of Deep Gravity models decreased with adding more layers. The 3-layer model of Deep Gravity performed the best, while deeper networks could not yield better results.

4.3.4 Risk Assessment of NIS through Ballast Water

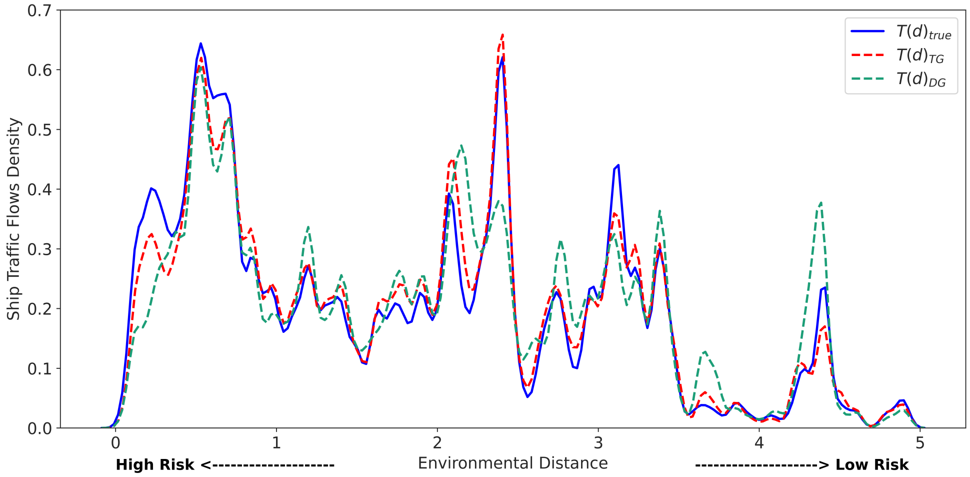


Figure 4.7: Distribution of environmental distances for shipping flows in 2019. The true shipping flows are represented by the blue curve, while the dashed red and green curves depict the predictions from the Transformer Gravity (TG) and Deep Gravity (DG) models, respectively. The x-axis measures the environmental distance; smaller values indicate higher risk levels, and larger values indicate lower risks.

Figure 4.7 shows the distributions of the three distance groups: $T(d)_{true}$, $T(d)_{TG}$, and $T(d)_{DG}$, from which we can find that $T(d)_{true}$ and $T(d)_{TG}$ are closely aligned in their trends, even in minor fluctuations. In contrast, $T(d)_{DG}$ is more differentiated from $T(d)_{true}$. To quantify the alignment between $T(d)_{TG}$ and $T(d)_{true}$, we calculated Pearson's correlation coefficients for the two groups of environmental distances, reaching the value of 0.889. This high coefficient indicates a strong linear correlation between the environmental distances associated with actual and predicted shipping flows. Figure 4.7 also reveals that the consistency between $T(d)_{TG}$ and $T(d)_{true}$ is more pronounced where the environmental distance is greater, implying a lower risk of invasion. Conversely, when smaller environmental distances indicate a higher invasion risk, the two groups show more discrepancies. Since the distance values in the two groups are calculated from the origin-destination pairs of the same year, these discrepancies can be attributed to the differences between the predicted and the real shipping flows. On the other hand, the Transformer Gravity model contains an important feature, *i.e.*, exportation values, quantifying the annual trading volume between two countries in US dollars. When contributing to the model's predictive performance, the feature value becomes zero and no longer positively associates trade with shipping flows when predicting inner-country shipping activities, and these are usually high-risk routes due to more environmental similarity between the source and destination locations. Further, there may also be critical factors in predicting high-risk shipping connections that remain unexplored, suggesting potential for further studies.

Although there are some discrepancies in the high-risk interval, the evaluation results show an overall low error of 0.208 and a high similarity between the environmental distances scaled according to the predicted and actual shipping flows. This result suggests that the high performance of our Transformer Gravity model for ship traffic flow prediction contributes significantly to representing shipping intensity in BWRA. This has contributed to a more accurate risk assessment for the spread of NIS through ballast water and thus provides a valuable reference for global ballast water risk management in the future.

4.4 Discussion

The ability to assess and predict the risks of NIS invasion through ballast water relies heavily on shipping fluxes. In order to improve BWRA, we have introduced a new approach called the Transformer Gravity model. Our Transformer Gravity model is built based on a stack of multi-head attention blocks. It incorporates features originating from the gravity-informed model, including the shipping fluxes, origin and destination locations, and the geographical distance between them. We enriched the feature set by integrating additional factors, including the international bilateral trade information and the graph metrics analyzed from the global shipping network, to enhance the model's capability for predicting ship traffic flows. We have validated our approach through a comprehensive comparison with established gravity models and a range of machine-learning regression techniques to highlight its advancements.

Our findings established the superiority of the Transformer Gravity model across several performance indicators. Notably, the model achieved the highest CPC with minimal variance across different data folds, indicating its robustness and adaptability to new datasets. The best average CPC in the cross-validation set was 0.864 for the 3 and 5-layer models. Besides, the 3-layer one demonstrated the lowest mean error, 0.080, and the highest mean correlation, 0.977. Variations in performance among baseline models highlighted their inadequacies in dealing with diverse data scenarios, thus reinforcing our model's resilience.

Our study has limitations that can be improved with future and further research. For instance, our model tends to be over-optimistic in targeting high-risk ship traffic predictions, and there might be underlying features that can contribute more to forecasting high-risk trips. Seasonality of environmental variables can affect the results of environmental distance calculation, and using environmental data collected from a smaller temporal scale can reflect the evaluated risk more precisely. Our study did not consider ecological barriers that divide ecoregions, which can overestimate the risk of some short-distance trips within the same ecoregions. Also, the change in shipping patterns due to COVID-19 requires further investigation into changes in shipping behavior and their impact on BWRA, which is a concern for future studies.

In summary, the Transformer Gravity model proposed in this study has markedly improved the performance of ship traffic flow forecast, outperforming other gravity-informed models. It has also shown high-performance consistency when applied to different data subsets with various spatio-temporal distributions. Future mobility studies and applications can enhance the model's explainability and transparency by exploring the intrinsic relationships between various features and the flow prediction results. The exploration of such aspects can improve the model's interpretability and help the model's application in domains where a high degree of accuracy is expected on the OD flow representation.

Acknowledgement

We thank all funding sources and *Spire* for providing the vessel trajectory dataset. We also thank Sarah Bailey, from the Great Lakes Laboratory for Fisheries and Aquatic Sciences, Fisheries and Oceans, Dartmouth – NS, Canada, for reviewing and sharing her thoughts about this paper.

Co-Authorship Statement

Conceptualization, R.S. and G.S.; methodology, R.S., G.S. and A.S.; validation, R.S., G.S., and A.S.; formal analysis, R.S., and G.S.; data modeling, R.S., and G.S.; data curation, R.S., G.S., and A.S.; writing - original draft preparation, R.S. and G.S.; writing - review and editing, all authors; visualization, R.S.; supervision, R.P., S.M., and A.S.; project administration, G.S. and A.S.; funding acquisition, R.P., S.M., A.S.;

Appendix: Concepts in Graph Analysis and Feature Information

Connected components: A component is a group of vertices that are connected to each other, and a network that has more than one group of vertices that are not connected is called a non-connected graph. In an arbitrary graph, vertices i and j are in the same component $G' = \{V', E'\}$, $V' \subseteq V$, if and only if $\{\forall i \in V', \forall j \in V' | d_{ij} < \infty\}$, meaning that it is possible to travel from any vertex i to any vertex j in a finite number of steps, where $d_{ij} : V \times V \rightarrow \mathbb{R}$ is a function that returns the distance between any two vertices [128]. Connected components are defined for undirected graphs but can also be extended to directed graphs, resulting in weekly and strongly connected components, which help in identifying absorbing regions.

Geodesic distance (d^E_{ij})***:*** The geodesic distance, as defined in Equation 4.12, measures the shortest distance between two vertices $i \in V$ and $j \in V$ on a sphere's surface [129]. This metric incorporates the latitudes ϕ_i and ϕ_j of vertices i and j, the difference $\triangle^{\lambda}_{ij}$ between their longitudes λ_i and λ_j, and the Earth's radius $\mathbf{R}$ (6,371 km). All values are calculated in radians.

$$d^E_{ij} = \left(sin\left(\phi_i\right) \times sin\left(\phi_j\right) + cos\left(\phi_i\right) \times cos\left(\phi_j\right) \times cos\left(\triangle^{\lambda}_{ij}\right) \times \mathbf{R}\right) \tag{4.12}$$

Shortest distance (d^N_{ij})***:*** A path S between two vertices i and j is a sequence of connected vertices $S^n = \langle v_1, v_2, ..., v_{q-1}, v_q \rangle$, where each consecutive vertex is connected through an edge $\left\langle S^n_m, S^n_{m+1} \right\rangle \in E$ for all $m \in [1, |S^n|[$. The shortest directed path d^N_{ij} is obtained by minimizing the weight function $f : E^n \rightarrow \mathbb{R}$, which describes the cost of the paths among all possible paths $\mathbb{S} = \{S^1, S^2, \ldots, S^n\}$ between vertices i and j [130]. The goal is to find the path with the minimum cost, which is determined by the sum of the weights of the edges. The weight is defined as the straight-line distance between the vertices, such as in Equation 4.13.

$$d_{ij}^{N} = min\left(\sum_{m=1}^{|S|-1} f(\langle S_m^n, S_{m+1}^n\rangle), \forall S^n \in \mathbb{S}\right) \tag{4.13}$$

Closeness centrality (C_C)***:*** The Closeness measures how closely a vertex is connected to all the other vertices in a network, as determined by the shortest paths between them (refer to Equation 4.14). A vertex with higher centrality is considered more central and has a shorter average distance to all other vertices in the network [131]. In a port network, the closeness-central ports are expected to have higher traffic of vessels, providing insight into trading behavior and economic relationships inherent to their country and cities.

$$C_C(i) = \frac{|V|-1}{\sum_{j=1}^{|V|} d_{ij}^{N}},\ i \neq j \tag{4.14}$$

Straightness centrality (C_S)***:*** This metric measures the straightness of paths connecting vertices i and j. It does so by comparing the deviation of the geodesic distance d_{ij}^{E} and the shortest path distance d_{ij}^{N} that links them [132]. A high centrality value indicates the existence of connections with distances close to the geodesic one. When the two distances match, it is the optimal scenario for communication between vertices.

$$C_S(i) = \frac{1}{|V|-1}\sum_{j=1}^{|V|}\frac{d_{ij}^{E}}{d_{ij}^{N}},\ i \neq j \tag{4.15}$$

Betweenness centrality (C_B)***:*** The Betweenness identifies how often a particular vertex is present in the shortest paths of a graph. A higher metric value indicates that the vertex is present in a larger number of shortest paths. In Equation 4.16, $\sigma(u, v)$ represents the number of shortest paths between vertices u and v, and $\sigma(u, v|i)$ represents the number of those paths that pass through vertex i. It is worth noting that there is also the *Edge-Betweenness*, which assigns a Betweenness value to each edge in a graph. This variant considers $\sigma(u, v|i)$ as the number of shortest paths between vertices u and v that pass through edge i.

$$C_B(i) = \sum_{\langle u,v\rangle \in V} \frac{\sigma(u, v|i)}{\sigma(u, v)},\ u \neq v \tag{4.16}$$

PageRank (P_{ij})***:*** The PageRank is based on a mathematical model known as the stochastic Markovian process. This model defines a probability distribution over a set of states, where the probability of transitioning from one state to another is solely correlated with the state immediately preceding it:

$$P_{ij} = \mathbf{P}(X_{n+1} = j \mid X_n = i) \tag{4.17}$$

The algorithm assesses the significance of a vertex with respect to the significance of the vertices that are linked to it. This way, it measures the vertex contribution based on the number of outgoing edges each adjacent vertex has, ensuring the uniqueness of the edge. To determine vertex importance, the algorithm calculates the stationary transition probability matrix. The values obtained from this calculation indicate the significance of the vertices based on their access probability. Equation 4.18 illustrates the stationary matrix, where $\pi^{(n)}$ denotes the probability matrix at time n, and $\mathbf{P}$ is the transition probability matrix.

$$\pi^{(n-1)}\mathbf{P} = \pi^{(n)} \tag{4.18}$$

Features information: Features incorporated in this study are listed in Table 4.3, including shipping fluxes, distances, international trade volume, and the graph metrics computed from the global shipping network.

Feature name	**Feature information**
Original port fluxes	Shipping fluxes at source port
Destination region fluxes	Total shipping fluxes at the destination geographical region
Distance	Geodesic distance between the source port to the destination region center
Exportation volume	Exportation volume in US dollars from source to destination
Betweenness centrality	Betweenness centrality at the original port and the median betweenness centrality in the destination region
Closeness centrality	Closeness centrality at the original port and the median closeness centrality in the destination region
Page rank	Page rank at the original port and the median page rank in the destination region

Table 4.3: Detailed information of features used in Transformer Gravity and other gravity-informed models.

Chapter 5

Conclusion and Future Works

This work explored the risk assessment of non-indigenous species (NIS) spreading through ballast water by examining the temporal variability of environmental factors that influence the calculations of environmental distance in the risk assessment of ballast water (BWRA) (details in Chapter 3) and forecasting the flows of shipping traffic within the global shipping network that reflect the intensity of shipping activities related to the volume of ballast water transported, as presented in Chapter 4.

In our examination of the effect of temporal variability of environmental factors on BWRA [46] in Chapter 3, the plotted standard deviation of the monthly sea surface salinity and temperature shows that estuaries of large rivers are more vulnerable to salinity variations, and the north hemisphere experienced more seasonal variations on the sea temperature. For ballast water destined for Canadian waters, the monthly temperature and salinity of the water source and destination were gathered as inputs to the BWRA model. The monthly environmental distances evaluated by the BWRA model were then statistically compared with the traditional annual environmental distances. The quantitative results indicated that the use of more accurate monthly environmental distance calculations resulted in an overall increase in the risk evaluated; that is, the traditional annual environmental distance assessment underestimated the risk of NIS invasion through ballast water, at least for those destined for Canada. Furthermore, we explored the variation in the risk estimated monthly for unique port pairs grouped by ballast water source region. These results show a general trend that ballast water from the northern hemisphere is at higher risk in summer and fall than in the winter and spring and that ballast water from Europe and East Asia poses an overall higher risk to Canadian waters.

These findings stress that the adoption of fine-grained and temporally sensitive data in BWRA is in demand, as we statistically prove that the annual data in the traditional assessment pipeline has underestimated the risk of NIS invasion traveling through ballast water during shipping activities. By incorporating monthly environmental data into BWRA models, we can better estimate the risk of NIS invasions, and more accurate predictions can help policymakers develop more effective ballast water management strategies.

As ballast water discharge volume in BWRA is influenced by shipping activities, we developed a comprehensive pipeline to forecast ship traffic flow within the global shipping network [47], as presented in Chapter 4. This pipeline integrates graph analysis in the global shipping network spanning from 2017 to 2019 and link prediction to filter out the real shipping connections in the network and to inform the prediction of shipping traffic flow sizes. Then, we proposed the Transformer Gravity model to predict the shipping flow sizes based on the estimated real shipping connections. As a predictive model, Transformer Gravity has driven insights from the traditional gravity model's principles with transformer neural networks. Unlike traditional gravity models that primarily rely on shipping fluxes and distances, the feature set of Transformer Gravity was enriched with the bilateral trade volume in currencies and the graph metrics derived from the analysis of the global shipping network. To evaluate the performance of our proposed framework, we included multilayer perceptrons

(MLPs)–based Deep Gravity model and multiple machine learning regression models, utilizing a consistent feature group for a fair comparison. The result shows our proposed Transformer Gravity model outperforms the Deep Gravity by 13.2% and improved by approximately 50% over other machine learning models. Further, our model has the most stable high performance and minimum variance in the validation folds, indicating its potential to capture intrinsic patterns with different data distributions. Further explorations into the model's architecture revealed that the stacked transformer encoder layers can enhance the model's performance. Among the configurations, the 3-layered model yields the best performance. The shipping flow sizes predicted by the Transformer Gravity model were then used to scale the environmental distances of all the port pairs in the shipping network. Compared with the environmental distances scaled by the Deep Gravity model, the distribution of the distances by Transformer Gravity exhibited a higher correlation with actual shipping conditions, as evidenced by a Pearson's correlation coefficient of 0.889.

This study advances our understanding of global shipping patterns that support the BWRA with high predictive accuracy of ship traffic flow. Leveraging the Transformer Gravity model, we can achieve a more accurate prediction of ship traffic flow, which helps assess environmental risks associated with ballast water discharge volumes. Moreover, the Transformer Gravity framework has the potential for application in a broader range of mobility research areas beyond the scope of marine shipping traffic, such as urban traffic patterns, social network analysis, and the study of species migration.

There are also several points in this thesis that can be improved in future studies. The first concerns the variation of salinity in the temporal assessment of environmental factors. While monthly average environmental data were used for environmental distance calculations for risk evaluation, salinity levels can exhibit significant fluctuations within a single day, especially in estuarine regions. Future studies can consider the areas with salinity dynamics to achieve a more accurate modeling of environmental factors. Secondly, the analysis revealed that the Transformer Gravity model tends to underpredict shipping flows for high-risk routes, leading to discrepancies between the predicted data distribution and the actual conditions. This suggests that additional features relevant to high-risk shipping routes might exist that could enhance the model's predictive accuracy. Future research can further study and explore these features to align the model's predictive performance more closely with real-world conditions. Lastly, shipping data used to construct and analyze the global shipping network covers the period from 2017 to 2019 in this work and does not consider the effect caused by the COVID-19 pandemic on global shipping patterns. Given the significant impact of the pandemic, future studies may need to invest more effort in analyzing post-COVID-19 shipping patterns. Updating the data to reflect recent changes will enable the BWRA tools to obtain up-to-date flow forecasts, thus improving the NIS risk assessment in the future.

Bibliography

[1] National Geospatial-Intelligence Agency (NGA). World Port Index (Pub 150). `https://msi.nga.mil/Publications/WPI`, 2019.

[2] Canada's national biodiversity clearing house. Invasive non-native species. `https://biodivcanada.chm-cbd.net/ecosystem-status-trends-2010/invasive-non-native-species`. [Accessed: 2021-11-27].

[3] Philip E. Hulme. Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, 46(1):10–18, 2009.

[4] Hanno Seebens, Tim M. Blackburn, Ellie E. Dyer, Piero Genovesi, Philip E. Hulme, Jonathan M. Jeschke, Shyama Pagad, Petr Pyšek, Marten Winter, Margarita Arianoutsou, Sven Bacher, Bernd Blasius, Giuseppe Brundu, César Capinha, Laura Celesti-Grapow, Wayne Dawson, Stefan Dullinger, Nicol Fuentes, Heinke Jäger, John Kartesz, Marc Kenis, Holger Kreft, Ingolf Kühn, Bernd Lenzner, Andrew Liebhold, Alexander Mosena, Dietmar Moser, Misako Nishino, David Pearman, Jan Pergl, Wolfgang Rabitsch, Julissa Rojas-Sandoval, Alain Roques, Stephanie Rorke, Silvia Rossinelli, Helen E. Roy, Riccardo Scalera, Stefan Schindler, Kateřina Štajerová, Barbara Tokarska-Guzik, Mark van Kleunen, Kevin Walker, Patrick Weigelt, Takehiko Yamanaka, and Franz Essl. No saturation in the accumulation of alien species worldwide. *Nature Communications*, 8(1):14435, 2017.

[5] Melodie A. McGeoch, Stuart H. M. Butchart, Dian Spear, Elrike Marais, Elizabeth J. Kleynhans, Andy Symes, Janice Chanson, and Michael Hoffmann. Global indicators of biological invasion: species numbers, biodiversity impact and policy responses. *Diversity and Distributions*, 16(1):95–108, 2010.

[6] Petr Pyšek, Philip E. Hulme, Dan Simberloff, Sven Bacher, Tim M. Blackburn, James T. Carlton, Wayne Dawson, Franz Essl, Llewellyn C. Foxcroft, Piero Genovesi, Jonathan M. Jeschke, Ingolf Kühn, Andrew M. Liebhold, Nicholas E. Mandrak, Laura A. Meyerson, Aníbal Pauchard, Jan Pergl, Helen E. Roy, Hanno Seebens, Mark van Kleunen, Montserrat Vilà, Michael J. Wingfield, and David M. Richardson. Scientists' warning on invasive alien species. *Biological Reviews*, 95(6):1511–1534, 2020.

[7] Reuben P Keller, Juergen Geist, Jonathan M Jeschke, and Ingolf Kühn. Invasive species in Europe: ecology, status, and policy. *Environmental Sciences Europe*, 23(1):1–17, 2011.

[8] Stephan Gollasch and Erkki Leppäkoski. *Initial risk assessment of alien species in Nordic coastal waters.* Nordic Council of Ministers, 1999.

[9] C. Clarke, Global Environment Facility, International Maritime Organization, Global Ballast Water Management Programme, and United Nations Development Programme. *Ballast Water Risk Assessment: Port of Khark Island, Islamic Republic of Iran : Final Report.* GloBallast Monograph Series. Programme Coordination Unit, Global Ballast Water Management Programme, 2003.

[10] A. Awad and Global Ballast Water Management Programme. *Ballast Water Risk Assessment: Port of Saldanha Bay Republic of South Africa : November 2003 : Final Report.* GloBallast Monograph Series. International Maritime Organization, 2004.

[11] Reuben P. Keller, John M. Drake, Mark B. Drew, and David M. Lodge. Linking environmental conditions and ship movements to estimate invasive species transport across the global shipping network. *Diversity and Distributions*, 17(1):93–102, January 2011.

[12] H. Seebens, M. T. Gastner, and B. Blasius. The risk of marine bioinvasion caused by global shipping. *Ecology Letters*, 16(6):782–790, June 2013.

[13] Matej David, Stephan Gollasch, and Erkki Leppäkoski. Risk assessment for exemptions from ballast water management – The Baltic Sea case study. *Marine Pollution Bulletin*, 75(1):205–217, 2013.

[14] Hanno Seebens, Nicole Schwartz, Peter J. Schupp, and Bernd Blasius. Predicting the spread of marine species introduced by global shipping. *Proceedings of the National Academy of Sciences*, 113(20):5646–5651, 2016.

[15] Mandana Saebi, Jian Xu, Erin K. Grey, David M. Lodge, James J. Corbett, and Nitesh Chawla. Higher-order patterns of aquatic species spread through the global shipping network. *PLOS ONE*, 15(7):1–24, July 2020.

[16] Mélanie Fournier, R. Casey Hilliard, Sara Rezaee, and Ronald Pelot. Past, present, and future of the satellite-based automatic identification system: Areas of applications (2004–2016). *WMU journal of maritime affairs*, 17(3):311–345, 2018.

[17] Amílcar Soares Júnior, Chiara Renso, and Stan Matwin. Analytic: An active learning system for trajectory classification. *IEEE computer graphics and applications*, 37(5):28–39, 2017.

[18] Pedram Adibi, Fabio Pranovi, Alessandra Raffaetà, Elisabetta Russo, Claudio Silvestri, Marta Simeoni, Amilcar Soares, and Stan Matwin. Predicting fishing effort and catch using semantic trajectories and machine learning. In *International Workshop on Multiple-Aspect Analysis of Semantic Trajectories*, pages 83–99. Springer International Publishing Cham, 2019.

[19] Amílcar Soares, Jordan Rose, Mohammad Etemad, Chiara Renso, and Stan Matwin. Vista: A visual analytics platform for semantic annotation of trajectories. In *Proceedings of the 22nd International Conference on Extending Database Technology (EDBT)*, 2019.

[20] Lucas May Petry, Amilcar Soares, Vania Bogorny, Bruno Brandoli, and Stan Matwin. Challenges in vessel behavior and anomaly detection: From classical machine learning to deep learning. In *Advances in Artificial Intelligence: 33rd Canadian Conference on Artificial Intelligence, Canadian AI 2020, Ottawa, ON, Canada, May 13–15, 2020, Proceedings 33*, pages 401–407. Springer, 2020.

[21] Damião Ribeiro de Almeida, Cláudio de Souza Baptista, Fabio Gomes de Andrade, and Amilcar Soares. A survey on big data for trajectory analytics. *ISPRS International Journal of Geo-Information*, 9(2):88, 2020.

[22] Fernando HO Abreu, Amilcar Soares, Fernando V Paulovich, and Stan Matwin. A trajectory scoring tool for local anomaly detection in maritime traffic using visual analytics. *ISPRS International Journal of Geo-Information*, 10(6):412, 2021.

[23] Fernando Henrique Oliveira Abreu, Amilcar Soares, Fernando V Paulovich, and Stan Matwin. Local anomaly detection in maritime traffic using visual analytics. In *In EDBT/ICDT Workshops*, 2021.

[24] Yashar Tavakoli, Lourdes Peña-Castillo, and Amilcar Soares. A study on the geometric and kinematic descriptors of trajectories in the classification of ship types. *Sensors*, 22(15):5588, 2022.

[25] Bruno Brandoli, Alessandra Raffaetà, Marta Simeoni, Pedram Adibi, Fateha Khanam Bappee, Fabio Pranovi, Giulia Rovinelli, Elisabetta Russo, Claudio Silvestri, Amilcar Soares, et al. From multiple aspect trajectories to predictive analysis: a case study on fishing vessels in the northern Adriatic sea. *GeoInformatica*, 26(4):551–579, 2022.

[26] Martha Dais Ferreira, Jessica Campbell, Evan Purney, Amilcar Soares, and Stan Matwin. Assessing compression algorithms to improve the efficiency of clustering analysis on AIS vessel trajectories. *International Journal of Geographical Information Science*, 37(3):660–683, 2023.

[27] Salman Haidri, Yaksh J Haranwala, Vania Bogorny, Chiara Renso, Vinicius Prado da Fonseca, and Amilcar Soares. Ptrail—a python package for parallel trajectory data preprocessing. *SoftwareX*, 19:101176, 2022.

[28] Yaksh J Haranwala, Salman Haidri, Terrence S Tricco, Vinicius P da Fonseca, and Amilcar Soares. A dashboard tool for mobility data mining preprocessing tasks. In *2022 23rd IEEE International Conference on Mobile Data Management (MDM)*, pages 278–281. IEEE, 2022.

[29] Simon C. Barry, Keith R. Hayes, Chad L. Hewitt, Hanna L. Behrens, Egil Dragsund, and Siri M. Bakke. Ballast water risk assessment: Principles, processes, and methods. *ICES Journal of Marine Science*, 65(2):121–131, March 2008.

[30] N. Paldor and D. A. Anati. Seasonal variations of temperature and salinity in the Gulf of Elat (Aqaba). *Deep Sea Research Part A. Oceanographic Research Papers*, 26(6):661–672, June 1979.

[31] Jean-Rene Donguy and Gary Meyers. Seasonal variations of sea-surface salinity and temperature in the tropical Indian Ocean. *Deep Sea Research Part I: Oceanographic Research Papers*, 43(2):117–138, February 1996.

[32] Thierry Delcroix, Marie-Hélène Radenac, Sophie Cravatte, Gaël Alory, Lionel Gourdeau, Fabien Léger, Awnesh Singh, and David Varillon. Sea surface temperature and salinity seasonal changes in the western Solomon and Bismarck Seas. *Journal of Geophysical Research: Oceans*, 119(4):2642–2657, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2013JC009733.

[33] Piers K. Dunstan, Scott D. Foster, Edward King, James Risbey, Terence J. O'Kane, Didier Monselesan, Alistair J. Hobday, Jason R. Hartog, and Peter A. Thompson. Global patterns of change and variation in sea surface temperature and chlorophyll a. *Scientific Reports*, 8(1):14624, October 2018.

[34] Johanna N. Bradie and Sarah A. Bailey. A decision support tool to prioritize ballast water compliance monitoring by ranking risk of non-indigenous species establishment. *Journal of Applied Ecology*, 58(3):587–595, March 2021.

[35] George Kingsley Zipf. The P1 P2/D Hypothesis: On the Intercity Movement of Persons. *American Sociological Review*, 11(6):677–686, 1946.

[36] Woo-Sung Jung, Fengzhong Wang, and H. Eugene Stanley. Gravity model in the Korean highway. *Europhysics Letters*, 81(4):48005, January 2008.

[37] Peter AG Van Bergeijk and Steven Brakman. *The Gravity Model in International Trade: Advances and Applications*. Cambridge University Press, 2010.

[38] Paulo Cesar Ventura, Alberto Aleta, Francisco Aparecido Rodrigues, and Yamir Moreno. Modeling the effects of social distancing on the large-scale spreading of diseases. *Epidemics*, 38:100544, 2022.

[39] Andrew M. Kramer, J. Tomlin Pulliam, Laura W. Alexander, Andrew W. Park, Pejman Rohani, and John M. Drake. Spatial spread of the West Africa Ebola epidemic. *Royal Society Open Science*, 3(8):160294, 2016.

[40] Pablo Kaluza, Andrea Kölzsch, Michael T. Gastner, and Bernd Blasius. The complex network of global cargo ship movements. *Journal of The Royal Society Interface*, 7(48):1093–1103, January 2010.

[41] César Ducruet, Hidekazu Itoh, and Justin Berli. Urban gravity in the global container shipping network. *Journal of Transport Geography*, 85:102729, May 2020.

[42] Gabriel Spadon, Andre CPLF de Carvalho, Jose F Rodrigues-Jr, and Luiz GA Alves. Reconstructing commuters network using machine learning and urban indicators. *Scientific reports*, 9(1):11801, 2019.

[43] Filippo Simini, Gianni Barlacchi, Massimilano Luca, and Luca Pappalardo. A Deep Gravity model for mobility flows generation. *Nature Communications*, 12(1):6576, November 2021.

[44] The Growth Lab at Harvard University. International trade data (SITC, rev. 2), 2019.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[46] Ruixin Song, Yashar Tavakoli, Sarah A. Bailey, and Amilcar Soares. A temporal assessment of risk of non-indigenous species introduction by ballast water to Canadian coastal waters based on environmental similarity. *Biological Invasions*, 25(6):1991–2005, June 2023.

[47] Ruixin Song, Gabriel Spadon, Ronald Pelot, Stan Matwin, and Amilcar Soares. Gravity-informed deep learning framework for predicting ship traffic flow and invasion risk of non-indigenous species via ballast water discharge, 2024.

[48] Sarah A. Bailey, Lyndsay Brown, Marnie L. Campbell, João Canning-Clode, James T. Carlton, Nuno Castro, Paula Chainho, Farrah T. Chan, Joel C. Creed, Amelia Curd, John Darling, Paul Fofonoff, Bella S. Galil, Chad L. Hewitt, Graeme J. Inglis, Inti Keith, Nicholas E. Mandrak, Agnese Marchini,

Cynthia H. McKenzie, Anna Occhipinti-Ambrogi, Henn Ojaveer, Larissa M. Pires-Teixeira, Tamara B. Robinson, Gregory M. Ruiz, Kimberley Seaward, Evangelina Schwindt, Mikhail O. Son, Thomas W. Therriault, and Aibin Zhan. Trends in the detection of aquatic non-indigenous species across global marine, estuarine and freshwater ecosystems: A 50-year perspective. *Diversity and Distributions*, 26(12):1780–1797, 2020.

[49] C. VAN DEN Hoek. The distribution of benthic marine algae in relation to the temperature regulation of their life histories. *Biological Journal of the Linnean Society*, 18(2):81–144, 06 1982.

[50] Johanna N. Bradie, Adam Pietrobon, and Brian Leung. Beyond species-specific assessments: an analysis and validation of environmental distance metrics for non-indigenous species risk assessment. *Biological invasions*, 17(12):3455–3465, 2015.

[51] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[52] Yaksh J. Haranwala, Gabriel Spadon, Chiara Renso, and Amilcar Soares. A data augmentation algorithm for trajectory data. In *1st ACM SIGSPATIAL International Workshop on Methods for Enriched Mobility Data: Emerging issues and Ethical perspectives 2023 (EMODE '23)*, page 5, New York, NY, USA, 2023. ACM, New York, NY, USA.

[53] César Ducruet, Justin Berli, Giannis Spiliopoulos, and Dimitris Zissis. *Maritime Network Analysis: Connectivity and Spatial Distribution*, pages 299–317. Springer International Publishing, Cham, 2021.

[54] Emanuele Carlini, Vinicius Monteiro de Lira, Amilcar Soares, Mohammad Etemad, Bruno Brandoli, and Stan Matwin. Understanding evolution of maritime networks from automatic identification system data. *GeoInformatica*, 26(3):479–503, July 2022.

[55] Zhenfu Li, Mengqiao Xu, and Yanlei Shi. Centrality in global shipping network basing on worldwide shipping areas. *GeoJournal*, 80(1):47–60, 2015.

[56] L. Jiang, L. Chen, W. Wang, W. Wei, Z. Lv, and H. Wang. Advanced Network Representation Learning for Container Shipping Network Analysis. *IEEE Network*, 35(2):182–187, March/April 2021.

[57] Nicanor García Álvarez, Belarmino Adenso-Díaz, and Laura Calzada-Infante. Maritime Traffic as a Complex Network: A Systematic Review. *Networks and Spatial Economics*, 21(2):387–417, June 2021.

[58] William L. Hamilton. *Graph Representation Learning*, volume 14. Morgan & Claypool Publishers, 2020.

[59] Linton C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1):35–41, 1977.

[60] Alex Bavelas. Communication Patterns in Task-Oriented Groups. *The Journal of the Acoustical Society of America*, 22(6):725–730, November 1950.

[61] Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31(4):581–603, December 1966.

[62] Kam-Fung Cheung, Michael G. H. Bell, Jing-Jing Pan, and Supun Perera. An eigenvector centrality analysis of world container shipping network connectivity. *Transportation Research Part E: Logistics and Transportation Review*, 140:101991, August 2020.

[63] Lawrence Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking : Bringing Order to the Web. In *The Web Conference*, November 1999. `https://www.eecs.harvard.edu/~michaelm/CS222/pagerank.pdf`.

[64] CÉSAR DUCRUET and THEO NOTTEBOOM. The worldwide maritime network of container shipping: Spatial structure and regional dynamics. *Global Networks*, 12(3):395–423, July 2012.

[65] Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. A Survey on Deep Learning for Human Mobility. *ACM Computing Surveys*, 55(1):7:1–7:44, November 2021.

[66] Isaac Newton, I.Bernard Cohen, and Anne Whitman. Proposition 75, Theorem 35. In *The Principia: Mathematical Principles of Natural Philosophy, 3rd Edition (1726)*, page 956. University of California Press, 1999.

[67] Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[68] Filippo Simini, Marta C. González, Amos Maritan, and Albert-László Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, April 2012.

[69] Yihui Ren, Mária Ercsey-Ravasz, Pu Wang, Marta C. González, and Zoltán Toroczkai. Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. *Nature Communications*, 5(1):5347, November 2014.

[70] Mattia Mazzoli, Alex Molas, Aleix Bassolas, Maxime Lenormand, Pere Colet, and José J. Ramasco. Field theory for recurrent mobility. *Nature Communications*, 10(1):3895, August 2019.

[71] Xin Yao, Yong Gao, Di Zhu, Ed Manley, Jiaoe Wang, and Yu Liu. Spatial Origin-Destination Flow Imputation Using Graph Convolutional Networks. *IEEE Transactions on Intelligent Transportation Systems*, 22(12):7474–7484, December 2021.

[72] Tim M. Blackburn, Petr Pyšek, Sven Bacher, James T. Carlton, Richard P. Duncan, Vojtěch Jarošík, John R.U. Wilson, and David M. Richardson. A proposed unified framework for biological invasions. *Trends in Ecology and Evolution*, 26(7):333–339, 2011.

[73] Mark A. Davis and Ken Thompson. Eight ways to be a colonizer; two ways to be an invader: A proposed nomenclature scheme for invasion ecology. *Bulletin of the Ecological Society of America*, 81(3):226–230, 2000.

[74] Robert I. Colautti, Igor A. Grigorovich, and Hugh J. MacIsaac. Propagule pressure: A null model for biological invasions. *Biological Invasions*, 8(5):1023–1037, 2006.

[75] Elizabeta Briski, Farrah T Chan, John A Darling, Velda Lauringson, Hugh J MacIsaac, Aibin Zhan, and Sarah A Bailey. Beyond propagule pressure: importance of selection during the transport stage of biological invasions. *Frontiers in Ecology and the Environment*, 16(6):345–353, 2018.

[76] Petr Pyšek and David M. Richardson. Invasive species, environmental change and management, and health. *Annual Review of Environment and Resources*, 35(1):25–55, 2010.

[77] Henn Ojaveer, Bella S Galil, Stephan Gollasch, Agnese Marchini, Dan Minchin, Anna Occhipinti-Ambrogi, Sergej Olenin, et al. Identifying the top issues of marine invasive alien species in Europe. *Management of Biological Invasions*, 5(2):81–84, 2014.

[78] Henn Ojaveer, Bella S. Galil, Marnie L. Campbell, James T. Carlton, João Canning-Clode, Elizabeth J. Cook, Alisha D. Davidson, Chad L. Hewitt, Anders Jelmert, Agnese Marchini, Cynthia H. McKenzie, Dan Minchin, Anna Occhipinti-Ambrogi, Sergej Olenin, and Gregory Ruiz. Classification of non-indigenous species based on their impacts: Considerations for application in marine management. *PLOS Biology*, 13(4):e1002130–, 04 2015.

[79] Gregory M. Ruiz, James T. Carlton, Edwin D. Grosholz, and Anson H. Hines. Global invasions of marine and estuarine habitats by non-indigenous species: Mechanisms, extent, and consequences. *American Zoologist*, 37(6):621–632, 12 1997.

[80] H Elçiçek, A Parlak, and M Cakmakci. Effect of ballast water on marine and coastal ecology. *Journal of Selcuk University Natural and Applied Science*, (1):454–463, 2013.

[81] Matej David, Stephan Gollasch, Erkki Leppäkoski, and Chad Hewitt. *Risk Assessment in Ballast Water Management*, pages 133–169. Springer Netherlands, Dordrecht, 2015.

[82] Mohammad Etemad, Amilcar Soares, Paul Mudroch, Sarah A. Bailey, and Stan Matwin. Developing an advanced information system to support ballast water management. *Management of Biological Invasions*, 13:68–80, 2021.

[83] M.M Zweng, J.R. Reagan, D. Seidov, T.P. Boyer, R.A Locarnini, H.E. Garcia, A.V. Mishonov, O.K Baranova, K.W. Weathers, C.R. Paver, and I.V. Smolyar. World ocean atlas 2018, volume 2: Salinity. `https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/`, 2019.

[84] R.A. Locarnini, A.V. Mishonov, O.K. Baranova, T.P. Boyer, M.M. Zweng, H.E. Garcia, J.R. Reagan, D. Seidov, K.W. Weathers, C.R. Paver, and I.V. Smolyar. World ocean atlas 2018, volume 1: Temperature. `https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/`, 2019.

[85] Sarah A. Bailey, Johanna N. Bradie, Dawson Ogilvie, and Paul Mudroch. Global port environmental data used for environmental distance calculations. *Dryad, Dataset*, 2020.

[86] Frank Wilcoxon. *Individual Comparisons by Ranking Methods*, pages 196–202. Springer New York, New York, NY, 1992.

[87] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[88] Alboukadel Kassambara. *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*, 2021. R package version 0.7.0.

[89] Codes for the representation of names of countries and their subdivisions - Part 2: Country subdivision code. Standard, International Organization for Standardization, Geneva, Switzerland, August 2020. `https://www.iso.org/standard/72483.html`.

[90] John C. Warner, W. Rockwell Geyer, and James A. Lerczak. Numerical modeling of an estuary: A comprehensive skill assessment. *Journal of Geophysical Research: Oceans*, 110(C5), 2005.

[91] Stephan Gollasch. *Untersuchungen des Arteintrages durch den internationalen Schiffsverkehr unter besonderer Berücksichtigung nichtheimischer Arten.* University of Hamburg, 1996.

[92] R. W. Hilliard, S. Walker, S. Raaymakers, and Ports Corporation of Queensland. *Ballast Water Risk Assessment, 12 Queensland ports : stages 2 and 3A*

report : selection & environmental descriptions of overseas source ports. Ports Corporation of Queensland Queensland, 1997.

[93] Jane Elith and John R Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution and Systematics*, 40(1):677–697, 2009.

[94] Michael Kearney and Warren Porter. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology letters*, 12(4):334–350, 2009.

[95] Lauren B. Buckley, Mark C. Urban, Michael J. Angilletta, Lisa G. Crozier, Leslie J. Rissler, and Michael W. Sears. Can mechanism inform species' distribution models? *Ecology Letters*, 13(8):1041–1054, 2010.

[96] Lennert Tyberghein, Heroen Verbruggen, Klaas Pauly, Charles Troupin, Frederic Mineur, and Olivier De Clerck. Bio-oracle: a global environmental dataset for marine species distribution modelling. *Global Ecology and Biogeography*, 21(2):272–281, 2012.

[97] John W. Williams and Stephen T. Jackson. Novel climates, no-analog communities, and ecological surprises. *Frontiers in Ecology and the Environment*, 5(9):475–482, 2007.

[98] Mike P. Austin and Kimberly P. Van Niel. Improving species distribution models for climate change studies: variable selection and scale. *Journal of Biogeography*, 38(1):1–8, 2011.

[99] Melanie A. Harsch and Janneke HilleRisLambers. Climate warming and seasonal precipitation change interact to limit species distribution shifts across western North America. *PLOS ONE*, 11(7):1–17, 07 2016.

[100] Jonas J. Lembrechts, Ivan Nijs, and Jonathan Lenoir. Incorporating microclimate into species distribution models. *Ecography*, 42(7):1267–1279, 2019.

[101] Joe Hereford, Johanna Schmitt, and David D. Ackerly. The seasonal climate niche predicts phenology and distribution of an ephemeral annual plant, Mollugo verticillata. *Journal of Ecology*, 105(5):1323–1334, 2017.

[102] Elizabeta Briski, Sarah A. Bailey, and Hugh J. MacIsaac. Invertebrates and their dormant eggs transported in ballast sediments of ships arriving to the Canadian coasts and the Laurentian Great Lakes. *Limnology and Oceanography*, 56(5):1929–1939, 2011.

[103] Sarah A. Bailey. An overview of thirty years of research on ballast water as a vector for aquatic invasive species to freshwater and marine environments. *Aquatic Ecosystem Health & Management*, 18(3):261–268, 07 2015.

[104] Statista Research Group. Container shipping - statistics & facts. https://www.statista.com/topics/1367/container-shipping/#topicOverview, May 2022.

[105] Osvaldo E Sala, FIII Stuart Chapin, Juan J Armesto, Eric Berlow, Janine Bloomfield, Rodolfo Dirzo, Elisabeth Huber-Sanwald, Laura F Huenneke, Robert B Jackson, Ann Kinzig, et al. Global biodiversity scenarios for the year 2100. *science*, 287(5459):1770–1774, 2000.

[106] Stephan Gollasch and Errki Leppäkoski. *Initial risk assessment of alien species in nordic coastal waters*. Nordic Council of Ministers, Copenhagen, Denmark, 8 1999.

[107] Mohammad Etemad, Amilcar Soares, Paul Mudroch, Sarah A. Bailey, and Stan Matwin. Developing an advanced information system to support ballast water management. *Management of Biological Invasions*, 13(1):68–80, 2022.

[108] Smithsonian Environmental Research Center. National Ballast Information Clearinghouse (NBIC) Database. https://nbic.si.edu/database/. Accessed: 2023-09-30.

[109] Gabriel Spadon, Jay Kumar, Matthew Smith, Sarah Vela, Romina Gehrmann, Derek Eden, Joshua van Berkel, Amilcar Soares, Ronan Fablet, Ronald Pelot, and Stan Matwin. Building a safer maritime environment through multi-path long-term vessel trajectory forecasting. *arXiv*, 2023.

[110] Gabriel Spadon, Martha D. Ferreira, Amilcar Soares, and Stan Matwin. Unfolding AIS transmission behavior for vessel movement modeling on noisy data leveraging machine learning. *IEEE Access*, 11:18821–18837, 2023.

[111] Duong Nguyen, Rodolphe Vadaine, Guillaume Hajduch, René Garello, and Ronan Fablet. Geotracknet - A maritime anomaly detector using probabilistic neural network representation of AIS tracks and A contrario detection. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5655–5667, 2022.

[112] Duong Nguyen and Ronan Fablet. TrAISformer-a generative transformer for AIS trajectory prediction. *CoRR*, abs/2109.03958, 2021.

[113] Martha Dais Ferreira, Gabriel Spadon, Amilcar Soares, and Stan Matwin. A semi-supervised methodology for fishing activity detection using the geometry behind the trajectory of multiple vessels. *Sensors*, 22(16):6063, 2022.

[114] Robert M. Beyer, Jacob Schewe, and Hermann Lotze-Campen. Gravity models do not explain, and cannot predict, international migration dynamics. *Humanities and Social Sciences Communications*, 9(1):1–10, February 2022.

[115] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, February 2010.

[116] A. Paolo Masucci, Joan Serras, Anders Johansson, and Michael Batty. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E: Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 88(2):022812, August 2013.

[117] J. Wang, J. Ji, Z. Jiang, and L. Sun. Traffic Flow Prediction Based on Spatiotemporal Potential Energy Fields. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14, 2022.

[118] Jose F Rodrigues-Jr, Gabriel Spadon, Bruno Brandoli, and Sihem Amer-Yahia. Patient trajectory prediction in the Mimic-III dataset, challenges and pitfalls. *CoRR*, abs/1909.04605, 2019.

[119] Jose F. Rodrigues-Jr, Marco A. Gutierrez, Gabriel Spadon, Bruno Brandoli, and Sihem Amer-Yahia. Lig-doctor: Efficient patient trajectory prediction using bidirectional minimal gated-recurrent networks. *Information Sciences*, 545:813–827, 2021.

[120] Gabriel Spadon, Shenda Hong, Bruno Brandoli, Stan Matwin, Jose F. Rodrigues-Jr, and Jimeng Sun. Pay attention to evolution: Time series forecasting with deep graph-evolution learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5368–5384, 2022.

[121] Bruno Brandoli, André R. de Geus, Jefferson R. Souza, Gabriel Spadon, Amilcar Soares, Jose F. Rodrigues, Jerzy Komorowski, and Stan Matwin. Aircraft fuselage corrosion detection using artificial intelligence. *Sensors*, 21(12), 2021.

[122] Gent Halili. Searoute-py: A python package to calculate the shortest sea route between two points on Earth. `https://github.com/genthalili/searoute-py`, 2022.

[123] Paolo Crucitti, Vito Latora, and Sergio Porta. Centrality measures in spatial networks of urban streets. *Phys. Rev. E*, 73:036125, Mar 2006.

[124] Ningwen Tu, Dimas Adiputranto, Xiaowen Fu, and Zhi-Chun Li. Shipping network design in a growth market: The case of Indonesia. *Special Issue on China's Belt and Road Initiative*, 117:108–125, September 2018.

[125] Maxime Lenormand, Aleix Bassolas, and José J. Ramasco. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography*, 51:158–169, 2016.

[126] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J. Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, March 2018.

[127] Sarah A Bailey, Johanna N. Bradie, Dawson Ogilvie, and Paul Mudroch. Global port environmental data used for environmental distance calculations [Dataset], December 2020.

[128] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein. *Introduction To Algorithms*. MIT Press, 2001.

[129] Takis Konstantopoulos. *Introduction to projective geometry*. Number September. Dover Publications, 2012.

[130] Maarten Van Steen. *Graph Theory and Complex Networks: An Introduction*. Maarten van Steen, 2010.

[131] Miray Kas, Kathleen M. Carley, and L. Richard Carley. Incremental Closeness Centrality for Dynamically Changing Social Networks. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM'13*. Association for Computing Machinery (ACM), 2013.

[132] I. Vragović, E. Louis, and A. Díaz-Guilera. Efficiency of informational transfer in regular and complex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 71:036122, Mar 2005.