ImPORTance! – Machine Learning-Driven Analysis of Global Port Significance and Network Dynamics for Improved Operational Efficiency

Emanuele Carlini
emanuele.carlini@isti.cnr.it
Inst. of Info. Sci. and Technologies
ISTI-CNR — Pisa, Italy

Domenico Di Gangi
digangidomenico@gmail.com
Inst. of Info. Sci. and Technologies
ISTI-CNR — Pisa, Italy

Vinicius Monteiro de Lira
vinicius.monteiro@insightlab.ufc.br
Federal University of Ceará
Fortaleza – CE, Brazil

Hanna Kavalionak
hanna.kavalionak@isti.cnr.it
Inst. of Info. Sci. and Technologies
ISTI-CNR — Pisa, Italy

Amilcar Soares
amilcar.soares@lnu.se
Linnaeus University
Växjö, Sweden

Gabriel Spadon*
spadon@dal.ca
Dalhousie University
Halifax – NS, Canada

Abstract

Seaports play a crucial role in the global economy, and researchers have sought to understand their significance through various studies. In this paper, we aim to explore the common characteristics shared by important ports by analyzing the network of connections formed by vessel movement among them. To accomplish this task, we adopt a bottom-up network construction approach that combines three years' worth of AIS (Automatic Identification System) data from around the world, constructing a Ports Network that represents the connections between different ports. Through this representation, we utilize machine learning to assess the relative significance of various port features. Our model examined such features and revealed that geographical characteristics and the port's depth are indicators of a port's importance to the Ports Network. Accordingly, this study employs a data-driven approach and utilizes machine learning to provide a comprehensive understanding of the factors contributing to the extent of ports. Our work aims to inform decision-making processes related to port development, resource allocation, and infrastructure planning within the industry.

CCS Concepts

• **Information systems** → **Location based services**; • **Computing methodologies**; • **Applied computing** → **Transportation**;

Keywords

AIS, Ports Network, Port Centrality, Port Importance, Connectivity

ACM Reference Format:
Emanuele Carlini, Domenico Di Gangi, Vinicius Monteiro de Lira, Hanna Kavalionak, Amilcar Soares, and Gabriel Spadon. 2025. ***ImPORTance***! – Machine Learning-Driven Analysis of Global Port Significance and Network Dynamics for Improved Operational Efficiency . In *19th International Symposium on Spatial and Temporal Data (SSTD '25), August 25–27, 2025, Osaka, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3748777.3748792

*Corresponding Author

SSTD '25, Osaka, Japan

ACM ISBN 979-8-4007-2094-9/25/08
https://doi.org/10.1145/3748777.3748792

1 Introduction

Seaports have garnered interest across economics, environmental studies, and social sciences. Understanding the role and significance of seaports (or simply *ports*) is crucial for analyzing ocean mobility, identifying major hubs, and assessing the capabilities of these ports.

A standard method for assessing port relevance is analyzing vessel routes and port interconnections, often represented as a Ports Network, where ports are nodes and interconnections are edges. These edges can denote sequences of port visits by vessels [3] or other maritime locations [27, 28]. Modeling ports as networks allows the use of graph theory to analyze their roles and structure [2, 7, 9, 11], offering a topological perspective [32]. A central outcome of using network analysis on top of Ports Networks is understanding port centrality [5, 12, 29], which reflects a port's topological importance.

As noted by Laxe et al.[11], centrality captures a port's geographic significance within maritime routes. These measures include static properties (*e.g.*, degree centrality) and flow-based metrics (*e.g.*, Betweenness, Closeness, PageRank). Many studies link network centrality to port relevance [18], although few explore the shared traits among worldwide ports. Historically, port network data came from shipping logs and insurance registries [9, 29]. More recently, Automatic Identification System (AIS) data have enabled detailed, large-scale Port Network modeling [3, 21, 30, 31], allowing precise vessel tracking and identification. AIS data also supports vessel traffic analysis and conservation planning for marine species [1, 15, 22].

In this paper, we construct a Ports Network from three years of global AIS data and examine how 34 features from the World Port Index[1] can predict port centrality. Our goals are: (i) to assess how well port characteristics reflect their importance, and (ii) to identify standard features among central ports. To this end, we train an AI-based model to predict port centrality using these features and analyze their importance with SAGE [6] and SHAP [14].

[1] https://msi.nga.mil/Publications/WPI

The feature importance analysis highlights that geographical location and port depth have the most influence in determining centrality. Building on these findings, our contributions are:

- The design of a centrality measure that combines multiple definitions of centrality used in literature and its application to a Ports Network built with three years of AIS data;
- The application and evaluation of a machine learning classification task to approximate port centrality in the network from the features extracted from their structure; and,
- An extensive feature importance analysis to evaluate relevant features of ports in estimating centrality.

In order to present these contributions, we structured this paper as follows: Section 2 reviews prior work on port centrality. Section 3 details the dataset, Ports Network construction, centrality definition, and machine learning methods. Section 4 presents key findings and feature importance results. Section 5 concludes the paper.

2 Related Works

We analyzed and compared the most relevant works in literature along three dimensions stated below (see Table 1 for more details):

- *Source data.* Ports Networks are usually built from two types of datasets: **(i)** bottom-up approaches use fine-grained data, such as AIS data; here, the connection between ports and vessel routes are extracted with an extensive data processing, such as in [31]; **(ii)** top-down approaches use ship schedules and historical registries of vessels routes to build the network. *We used three years (2017-2019) to build a Ports Network.*
- *Network scope.* Studies vary regarding the location of the ports. Analysis can be focused on a specific area or country, such as the Canary Islands [25], or global trends (*i.e.*, worldwide and continental). *We have analyzed all the world's ports.*
- *Centrality measures.* A wide variety of centrality measures have been used in the literature. Our analysis derives an aggregated centrality measure combining the Degree, PageRank, Betweenness, and Closeness Centralities. Proposing a new technique focused on the importance of nodes, and with a specific aim of assessing ports in such a Port Network.

Ducruet et al. [8] are among the first to systematically study cargo networks with complex network techniques at a large scale. In one of their first papers [8], they provide an empirical analysis of the centrality of ports in Northeast Asia by using inter-port traffic flows from the Lloyd's Shipping Index[2]. The work of [11] also utilizes a sample of the Lloyd's dataset, which includes the movements of the world's container ship fleet from Chinese ports from 2008 to 2010. Their work aims to examine the maritime networks before and after the 2008 financial crisis, analyzing the extent to which large ports have seen their position within the network change. The authors demonstrate how the global and local significance of ports can be quantified using concepts from graphs.

A similar analysis is conducted by [16], which uses the same centrality metrics to observe the evolution of the Ports Network of container vessels on a broader geographical scale for 2 months of 3 consecutive years. A complete historical analysis of the Lloyds dataset was conducted by [10]. They investigate the structure of the maritime trade network and examine its relationship to efficiency. Given the extended coverage of their data (1890-2000), the authors assessed the topological over-time change of the corresponding network, which evolved from a highly clustered network towards a hub-and-spoke network. Cheung et al. [5] utilized eigenvector centrality as a decision-making tool for predicting potential new links between ports that would enhance network connectivity. In other words, the eigenvector centrality becomes the objective of a max-min optimization problem. They built the network using information from major shipping lines in the fourth quarter of 2015, resulting in a graph comprising 601 nodes (ports) and 3,737 links.

The work of [19] employs complex network techniques to analyze the growth rates of European ports. The "proximal foreland" (essentially a subgraph of a Ports Network considering all ports at three hops of a given center port) is used to measure the connectivity of the ports. They analyzed the AIS data of one month of four different years (2007, 2008, 2010, 2011) to extrapolate cargo vessel trajectories. AIS data were also utilized by [30], who constructed a Ports Network using the 2015 worldwide AIS data with multiple spatial levels. Their bottom-up process mainly consists of five steps, where the first three generate the network nodes, and the last two create the links. They apply the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to detect where ships stop and cross, and then combine this information with terminal candidates of ports. A directed Port Network is generated with the trip statistics between two nodes as the edges. They evaluate features such as the average Degree and Betweenness node

	Data Source (Period)	Network Scope	Centrality
[8]	LSI (1996-2006)	Local (NE Asia)	MD, B
[11]	LSI (2008-2010)	Local (China)	B
[16]	LSI (2008-2010)	Global (World)	B
[19]	AIS (Mar 2007-08, 2010-11)	Local (Europe)	B, D
[26]	CIY (1995-2011)	Global (E-W lane)	D*, WD*
[25]	SLS (Oct 2012)	Local (Canary Island)	D, B, AI
[10]	LSI (1890-2000)	Global (World)	C, B
[30]	AIS (2015)	Global (World)	D, B
[5]	SLS (Q4 2015)	Global (World)	E
[24]	SLS (2019)	Global (World)	B*, D
[29]	SLS (N/S)	Global (39 ports)	D, C, B
[12]	AIS (2017)	Regional (Chesapeake Bay)	PR, PC
Ours	AIS (2017-2019)	Global (World)	D, PR, B, C

Table 1: Comparison of relevant works in centrality measurement on port networks — the asterisk "*" indicates an adapted version of the centrality with respect to the original formulation in classical graph theory; *Acronyms:* LSI – Lloyd's Shipping Index, AIS – Automatic Identification System, CIY – Containerisation International Yearbooks, SLS – Shipping lines schedule, MD – Maritime Degree, B – Betweenness, D – Degree, WD – Weighted Degree, C – Closeness, E – Eigenvector, PR – PageRank, PC – Participation Coefficient, AI – Accessibility Index, N/S – Not Specified.

[2]The insurance company Lloyd's has historically collected movement of vessels between two or more ports daily or weekly since 1890. Their data includes the dates of departure and arrival, tonnage capacity, operating company, flag, and additional information on the voyage. Nowadays, Lloyd's covers about 80% of the world's fleet.

centrality, the average shortest path length between any two nodes, and community clusters within their proposed Ports Network.

Several works employ a modified version of the original centrality metrics to adapt for the peculiarity of Port Networks. The work presented in [26] analyses the East-West lane over several years. They compute a set of indicators to evaluate the evolution of connectivity in terms of degree centrality, which is also calculated by considering the amount of TEU (twenty-foot equivalent units) exchanged between ports. The work of [24] focuses on applying centrality measures on a hypergraph network of container services. They provide centrality for (hyper) P-graphs, which represent direct port-to-port service connections, and (hyper) L-graphs, where the edges represent the transit of a container vessel between two ports. For reference, our paper considers the network as an L-graph. The Betweenness Centrality for hypergraphs is computed using the probability that the shortest path goes through a specific node in the hypergraph (rather than the yes/no formulation of the classical Betweenness Centrality). In [29], the authors define the importance of a port by applying a *centrality index*, composed of Freeman's measures of Degree, Closeness, and Betweenness Centrality measures. They utilize these centrality measures to select 39 container ports worldwide (no information about the temporal range provided).

Network analysis has also been used to explore the relevance of ports localized in specific regions. [25] assesses the connectivity of the main Canarian ports with various centrality measures (*e.g.* Degree, Betweenness, and Port Accessibility Index). They generated a network of 53 ports directly related to Las Palmas and Tenerife ports, using one month (October 2012) of the shipping line schedule. More recently, the work of [12] used AIS data to generate trajectory sequences and assess the importance of way-points in the Chesapeake Bay area. They employed one year (2017) of AIS data to generate a fine-grained network of an enclosed area and compute a centrality analysis of way-points. To assess the importance of way-points, they use both centrality measures (PageRank) and community-based measures, *i.e.*, the participation coefficient (the strength of a node's connections within its community).

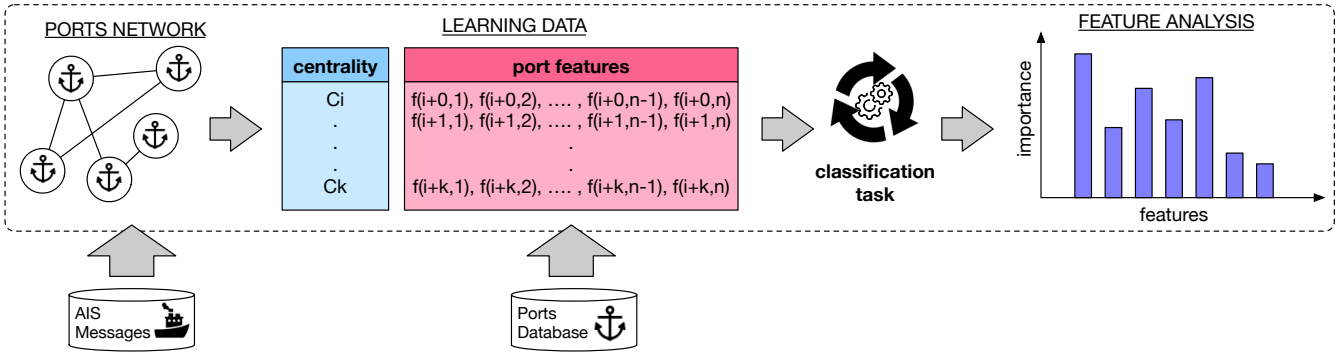

Figure 1: Methodological Framework for Port Centrality and Feature Analysis. This diagram outlines the methodology used to analyze port centrality and its features following the bottom-up approach to creating our Ports Network. Starting from the AIS messages and port databases, the Ports Network is constructed to calculate centrality measures. Because various features form the basis of the data used for the classification task, the feature analysis identifies key factors contributing to port centrality.

3 Dataset & Methodology

This paper investigates how port features relate to their importance in maritime networks of ports interconnectivity. Through our methodology we seek response for two queries: **(RQ1)** *How effective is it to assess port importance by solely considering its features?* and **(RQ2)** *What are the common features found in the most central ports?*

We implemented the methodology in Figure 1 to answer these questions. The Ports Network is built by using two datasets: **(i)** a publicly available dataset containing the port features and position, and **(ii)** an AIS messages dataset that contains information about vessels for the period 2017-2019. The importance of ports is then defined by considering their centrality in the network. The relationship between importance and features is assessed using a machine learning model trained to predict significance based on features.

3.1 Data Sources

AIS data. It is a key data source for tracking vessel movements, transmitting positions and activity data globally Each message contains several pieces of information regarding a vessel, including the ship identifiers, coordinates, and other data such as kinematics and vessel type. More details on the format and applications of AIS messages can be found in [31]. In this work, we used three years (2017-2019) of worldwide AIS data provided by Spire[3]. The raw dataset is private and contains 20 billion messages, around 2.5 TB.

Port data. Ports' features are taken from the 2020 World Port Index dataset[4] (WPI), a publicly available dataset released by the National Geospatial-Intelligence Agency of United States. The dataset contains information on about 3,630 ports worldwide. In total, 79 features are reported for each port. They include various port identifiers, positions, port characteristics, facilities, and available services.

3.2 Ports Network

Vessels report their positions (*i.e.*, coordinates) via AIS messages during navigation. We processed these messages to capture the

[3] spire.com (now operated by kpler.com)
[4] msi.nga.mil/Publications/WPI

fine-grained spatio-temporal trajectories of individual vessels. From these, we extracted sequences of port visits for each unique vessel, ordered by time. All vessel sequences were then merged to construct a unified Ports Network, where nodes represent ports and edges capture traffic between them. An edge is created when a vessel visits two ports consecutively, and its weight reflects the total number of vessels traveling between those ports. Construction the Ports Network involves three steps: **(i)** extracting port visits from raw AIS data; **(ii)** aggregating port-to-port transitions; and **(iii)** generating the network from the compiled visit sequences. These procedures are detailed in our earlier work [3, 4].

3.2.1 Port Visits. This phase aims to identify vessel visits to ports. First, we filter AIS messages transmitted within defined port areas – modeled as circular buffers centered on each port's coordinates. To handle overlapping buffers (*i.e.*, AIS messages that fall within multiple ports), we merge overlapping areas and retain the largest port, determined using the HARBORSIZE attribute. Vessels are uniquely identified by a combination of MMSI, IMO number, and call sign; entries lacking this information were excluded. Next, for each vessel, AIS messages are chronologically sorted to extract a sequence of port visits. Consecutive messages from the same port are collapsed into a single visit. This process yielded approximately 4.5 million visits by nearly 100,000 uniquely identified vessels.

3.2.2 Generating the Ports Network. We constructed the Ports Network using only *cargo* vessels, totaling approximately 1.5 million port visits from 30,000 unique vessels. From these visits, we derived each vessel's sequence of voyages, where a voyage is defined as the movement from an origin port p_o to a destination port p_d based on consecutive timestamps t_i and t_{i+1}. This data was used to build a directed graph $G = V, E$, where V denotes ports and E represents directed edges corresponding to vessel movements. Initially, this produced a multi-graph with one edge per voyage. We then collapsed multi-edges into single directed edges weighted by the frequency of voyages between port pairs. The resulting network consists of multiple strongly connected components. Our analysis focuses on the largest one, which includes 1,154 nodes (93%) and 21,776 edges (99%), of which 7,544 (35%) are bi-directional. The network is sparse, with a density of 0.01, a diameter of 10, an average shortest path length of 3.1, and a clustering coefficient of 0.6.

3.3 Port Centrality

Several centrality measures can be utilized to the assessment of the relevance of a port. To be as general as possible, we have chosen to measure centrality by combining the most used metrics in the literature we reviewed. Specifically, we define port centrality as the function $A(\cdot)$ defined for each port $p \in V$ in the Ports Network. $A(\cdot)$ compute the aggregation of the following centrality measures:

- *InDegree* (DI) and *OutDegree* (DO) are defined by counting, respectively, the number of incoming and outgoing links for each node. In this context, the DO of a port refers to the number of ports from which vessels depart, while the DI counts the number of ports to which vessels arrive.
- *PageRank* (PR) and its weighted variant (wPR) are widely used algorithms for evaluating node importance in complex networks. Initially developed to rank web pages, PR has since been applied across various domains, including social networks and biological systems. The algorithm assigns importance to a node based on the significance of its neighbors. In our context, a port's relevance is proportional to the importance of other ports connected to it via vessel movements. In the classical (unweighted) PR, each neighboring port contributes equally to the overall score. In contrast, the weighted version (wPR) adjusts contributions based on the logarithm of the number of trips between connected ports.
- *Betweenness Centrality* (BC). The BC is a topological measure of a node's importance. It is widely adopted for analyzing a wide range of complex networks, particularly those with a concept of information flow between nodes (in a Ports Network, the flow represents the traffic between ports). The BC works by counting the number of times a node is in the shortest path between all node pairs in the network, as in:

$$BC(p) = \sum_{p \neq s \neq d} \frac{\sigma_{sd}(p)}{\sigma_{sd}} \tag{1}$$

  where σ_{sd} is the number of shortest paths from a starting node s and a destination node d, and $\sigma_{sd}(p)$ is the number of shortest paths that go through p. In our case, the flow is represented by the maritime traffic between ports, and the importance of a port is proportional to the presence of the port in the critical routes of the global Ports Network.
- *Closeness Centrality* (CC). It measures the closeness of a node to all others. The CC of a node n is the reciprocal of the average shortest path distance from it to all reachable ones:

$$CC(n) = \frac{|N|}{\sum_{y \in N} d(y, n)} \tag{2}$$

  where N is the set of all nodes in the network; indicating the proximity of a port in terms of "hops" to other ports.

Applying each of those measures to the Ports Network produces one centrality measurement for each node. Therefore, for each port p we compute the 6 centralities described above resulting in:

$$\mathscr{C} = \{DI(p), DO(p), PR(p), wPR(p), BC(p), CC(p)\}. \tag{3}$$

To combine the different measures, we first standardize each by computing the z-score, which involves subtracting the mean and then dividing by the standard deviation. In the case of a generic centrality $c \in \mathscr{C}$ for a port p and N ports, formalized as:

$$z(c, p) = \frac{c(p) - \frac{\sum_p c(p)}{N}}{\sqrt{\sum_p c(p)^2 - \left(\sum_p c(p)\right)^2}} \tag{4}$$

Finally, the final measure of the aggregated centrality of a port is the average of the standardized centralities, formalized as:

$$A(p) = \frac{\sum_{c \in \mathscr{C}} z(c, p)}{|\mathscr{C}|} \tag{5}$$

3.4 Port Features

The WPI dataset (see Section 3.1) contains 70+ features for 3,630 ports. The preprocessing of the WPI dataset included cleaning operations and filling empty data. From the entire dataset, we removed those features that represent meta or external references, such as

Category	Attribute	Missing(%)	Support	Cardinality	Min Distr. (%)	Max Distr. (%)
COMMUNICATION	**AIR**	41.3	2130	2	3.4	96.6
	FAX	66.0	1234	2	2.6	97.4
	PHONE	47.2	1915	2	3.1	96.9
	RADIO	29.6	2556	2	1.1	98.9
	RADIO TEL	56.9	1563	2	3.0	97.0
	RAIL	50.6	1795	2	2.4	97.6
CRANE	FIXED	61.8	1387	2	6.9	93.1
	FLOAT	82.0	652	2	16.3	83.7
	MOBIL	54.3	1658	2	5.2	94.8
DEPTH	**ANCHORAGE**	9.8	3274	16	0.1	18.5
	CARGO DEPTH	12.5	3175	16	0.3	16.2
	CARGO WHARF	24.3	2747	2	0.3	99.7
	CHANNEL	12.6	3171	16	0.9	11.5
	OIL WHARF	0.0	1689	15	1.5	14.7
ENTRANCE RESTRICTIONS	**ICE**	26.1	2681	2	22.9	77.1
	OTHER	15.6	3064	2	7.9	92.1
	SWELL	22.1	2827	2	27.1	72.9
	TIDE	24.6	2736	2	29.2	70.8
LIFT	50-100 TONS	79.5	743	2	7.7	92.3
OTHER	CARGO_ANCH	56.7	1573	2	2.3	97.7
	DIRTY BALLAST	28.4	2600	2	34.7	65.3
	DRYDOCK	79.4	748	3	19.0	43.9
	ETA MESSAGE	15.9	3054	2	10.8	89.2
	FIRST PORT OF ENTRY	37.8	2259	2	24.3	75.7
	GARBAGE DISPOSAL	55.7	1608	2	20.9	79.1
	HARBOR SIZE	0.2	3624	4	4.4	58.6
	HARBOR TYPE	0.0	3618	8	0.9	34.9
	HOLDGROUND	47.7	1899	2	13.7	86.3
	MAX SIZE VESSEL	19.5	2921	2	39.1	60.9
	MEDICAL FACILITIES	23.5	2776	2	3.2	96.8
	OVERHEAD LIMITATION	48.5	1868	2	39.2	60.8
	RAILWAY	58.8	1497	3	9.6	65.4
	REPAIR_CODE	24.6	2737	5	5.2	60.7
	SHELTER	0.9	3599	5	1.0	35.7
	TIDE_RANGE	0.0	3587	14	0.5	32.4
	TURNING AREA	63.3	1332	2	11.4	88.6
	US REPRESENTATIVE	38.2	2244	2	14.4	85.6
PILOTAGE	ADVISABLE	64.5	1290	2	5.3	94.7
	AVAILABLE	30.5	2522	2	5.1	94.9
	REQD	19.9	2907	2	14.6	85.4
QUARENTINE	OTHER	69.2	1119	2	0.3	99.7
	PRATIQUE	45.0	1998	2	2.1	97.9
SERVICES	ELECTRICAL	72.5	999	2	12.1	87.9
	LONGSHORE	46.4	1946	2	5.8	94.2
SPATIAL	**COUNTRY**	0.0	3029	48	0.5	22.0
	LAT_HEMISPHERE	0.0	3629	2	13.8	86.2
	LONG_HEMISPHERE	0.0	3629	2	47.3	52.7
	LATITUDE	0.0	3629	cont.	n/a	n/a
	LONGITUDE	0.0	3629	cont.	n/a	n/a
SUPPLIES	DECK	63.8	1315	2	24.0	76.0
	DIESEL	38.6	2228	2	15.9	84.1
	ENGINE	63.7	1316	2	23.0	77.0
	FUEL OIL	26.0	2686	2	14.7	85.3
	PROVISIONS	37.9	2255	2	7.2	92.8
	WATER	13.6	3136	2	6.6	93.4
TUGS	**ASSIST**	26.2	2680	2	22.2	77.8
	SALVAGE	67.7	1172	2	23.5	76.5

Table 2: Summary of Key Port Features in the WPI Dataset. Features are grouped by category. Features retained after the cleaning phase are shown in bold. Missing indicates the percentage of null. Support is the number of ports with valid entries. Cardinality is the number of unique feature classes. Min and Max is the proportion of the least and most frequent classes.

the region number and the port name. From the remaining features, as listed in Table 2, we removed those with a percentage of missing values exceeding 50%. This results in 36 features (in **bold** in Table 2) for the analysis, of which 34 are categorical and 2 are continuous.

We then impute the remaining missing values using a multivariate iterative approach. That amounts to sequentially fitting a regression to explain the non-missing values of one column as a function of all the other columns as discussed, for example, in [13]. The regression is used to infer the missing values of the column, and this procedure is repeated for each column. The whole cycle of the regression fit and imputations is repeated 10 times[5].

[5]We used scikit-learn's iterative imputation: tinyurl.com/35b399hn.

3.5 Predicting Ports Relevance

To predict relevance from the features, we first consider a binary classifier to determine whether a port is among the most central ones. We experimented with thresholds to separate relevant and non-relevant ports, considering the top 5%, 10%, and 15% as relevant. We use 75% of the dataset for the training and 25% for evaluation.

We examined various machine learning models for the binary classification task and consistently obtained similar results across different models. We selected Random Forest to distinguish central from non-central ports effectively. The following section discusses the results obtained from this approach, which mitigates the high variance found in single trees by averaging their numbers. Each tree grows considering only a random subset of features before each split. After a fixed number of different trees are grown in this way, their predictions are combined with a majority vote. We used 100 trees, grown using Gini impurity to guide tree splits.

3.6 Evaluate Feature Importance

Explaining the outcomes of machine learning models is a significant problem in many contexts and has garnered considerable attention in the literature. Various interpretability tools and approaches have been proposed to improve our understanding of complex models [14, 17, 23]. Here we consider two approaches, SHapley Additive exPlanations (SHAP) [14] and Shapley Additive Global importancE (SAGE) [6], that quantify how much a model relies on each feature in making predictions and define feature importance as the amount of predictive power that can be associated to it. While both methods are inspired by game theory concepts and are based on Shapley values, they differ in the scope of the feature importance they define. On the one hand, SHAP explains the outcome of individual predictions, thus focusing on local interpretability. On the other hand, SAGE describes the model's behavior across the entire dataset, which we refer to as global interpretability. In the following, we provide an overview of the two methods and refer the interested reader to the original papers for more detailed information.

SHAP and SAGE are based on the Shapley values [20]. These methods enable the computation of each feature's importance while considering its interactions. To do so, they repeatedly evaluate the model with various input feature sets derived from the original one by "soft-removing" the information content of features not included in the set. The removal of a feature is achieved by shuffling its values, thereby destroying any potential contribution to the prediction. In the following, we indicate random variables with capital letters and observations in lowercase. Given the response variable Y, the set of features X, consisting of individual features $X_1, \ldots, X_d$, a subset S of the features is denoted by X_S. Considering a model $f(X)$, the methods consider a feature necessary if its absence harms the model outcome, where the model impact is quantified differently in SHAP and SAGE. Instead of considering the removal of a single feature, both approaches, based on Shapley values, compare all possible subsets of features, including or excluding each feature.

3.6.1 Local Interpretability. SHAP [14] captures the relevance of each feature for an individual prediction, rather than considering each feature's relevance for the entire set of predictions. In our application, SHAP answers questions like *how relevant was feature cargo depth in classifying Port Keppel as central (non-central)?*

SHAP starts from the prediction for one observation x[6], based features x_S. That is defined as $v_{f,x}(S) = \mathbb{E}[f(X)|X_S = x_S]$ where the expectation is taken over the rest of the features not in S. Those features are marginalized out, and their contribution is "softly-removed". From this the SHAP value for x $\phi_i(f,x)$ is defined as:

$$\phi_i(f,x) = \frac{1}{d} \sum_{S \in D \setminus i} \binom{d-1}{|S|} \left[v_{f,x}(S \cup \{i\}) - v_{f,x}(S) \right] \quad (6)$$

A negative value indicates that including a feature in the model decreases the expected prediction for the observation.

3.6.2 Global Interpretability. SAGE focuses on the global importance of a feature for all observations available. This amounts to assigning a number to each feature to quantify its relevance for predicting all observations. In our application, SAGE answers questions like *how relevant was feature cargo depth in classifying ports as central or non-central?* They consider a loss function $l(f(X), Y)$, *i.e.*, the cross entropy, and define $f_S(x_S) = \mathbb{E}[f(X)|X_S = x_S]$ where the expectation is taken over the rest of the features not in S. Then they define the amount of predictive power that a model derives from the set of features S as $v_f(S) = \mathbb{E}[l(f_\emptyset(X_\emptyset))] - \mathbb{E}[l(f_S(X_S))]$. To obtain the importance of individual feature i from v_S (a set S of possibly multiple features), they compare all the sets, including or not including feature i – the SAGE value ϕ_i is:

$$\phi_i = \frac{1}{d} \sum_{S \in D \bmod i} \binom{d-1}{|S|} \left[v(S \cup \{i\}) - v(S) \right] \quad (7)$$

A negative SAGE value indicates that a feature contributes negatively to the overall model performance, meaning that considering that feature is detrimental to the model.

4 Evaluation & Findings

4.1 Most central ports

Figure 2 shows the geographical placement of the most central ports (the size and darkness of the circle are proportional to the port centrality). Similarly, Table 3 shows the top 12 most central ports according to the definition of centrality given in Equation 5. Examining the list in the table, we recognize some major global container hubs. The top ports show a relatively wide geographical distribution. First in the top 12, *Keppel* in Singapore is located in one of the busiest port areas in the world. *Algeciras* in Spain and *Tangier* in Morocco are located on the Strait of Gibraltar, the Mediterranean's entrance passage, and a stopping point for cargo ships. *Puerto Cristobal* (Colon) is a ship hub on the Panama Canal's Atlantic entrance. *Las Palmas* is the largest port of the Canary Islands, and it's strategically positioned on the Atlantic Ocean.

Alexandria and *El-Adabiya* are two major Egyptian ports, respectively, on the Mediterranean and at the Suez Canal entrance in the Suez Gulf. *Yokohama* is one of the largest cities in Japan, situated in a highly congested area at the entrance to Tokyo Port. The *Durban* port in South Africa is one of the largest container ports in the Southern Hemisphere. *Jakarta* and *Gresik* ports are the two largest in Indonesia, with Jakarta being one of the largest in Southeast Asia. Finally, *Casablanca* is the second-largest city in Morocco.

[6] In our case, each observation is a port with its set of features.

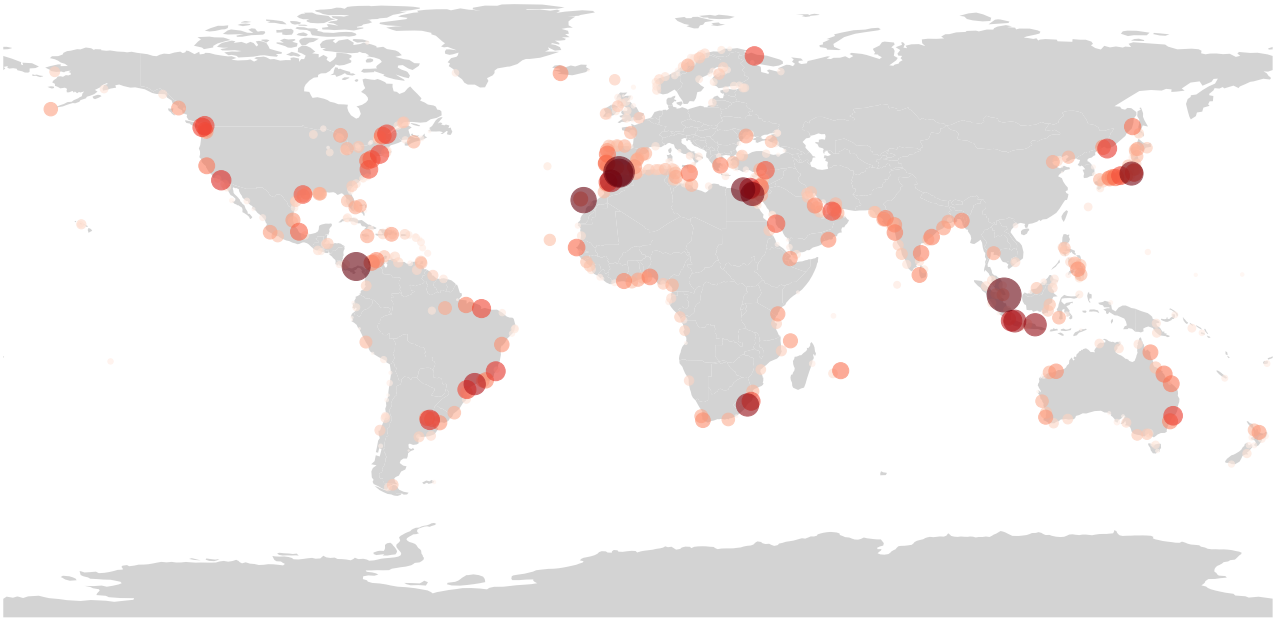

Figure 2: Spatial Distribution and Centrality of Global Maritime Ports. This map displays the locations of major ports worldwide. The size of each circle corresponds to the importance and connectivity of the port within the global maritime network. Larger circles represent ports with higher centrality, indicating their significance in international trade and logistics.

Rank	Port Name	Country	Centrality
1	KEPPEL	Singapore	10.594
2	ALGECIRAS	Spain	8.554
3	PUERTO CRISTOBAL	Panama	7.182
4	TANGIER	Morocco	6.742
5	LAS PALMAS	Spain	5.89
6	ALEXANDRIA	Egypt	4.941
7	YOKOHAMA	Japan	4.882
8	EL-ADABIYA	Egypt	4.823
9	GRESIK	Indonesia	4.542
10	DURBAN	South Africa	4.538
11	JAKARTA	Indonesia	4.4
12	CASABLANCA	Morocco	4.271

Table 3: Top 12 central ports and associated centrality values.

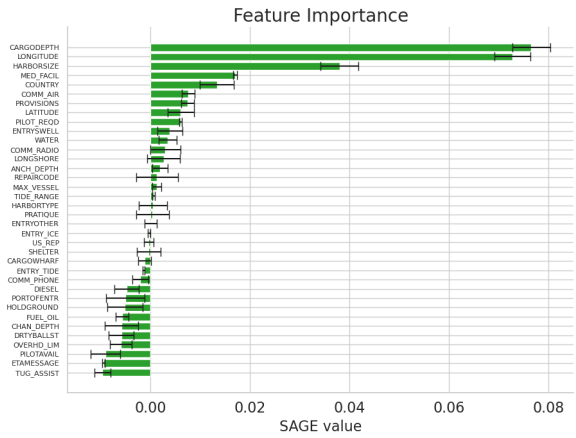

Figure 3: SAGE-Derived Feature Importance for Central Port Classification. This bar chart illustrates the global feature importance scores computed using the SAGE method for identifying the top 10% most central ports. Features such as CARGODEPTH, LONGITUDE, HARBORSIZE, and MED_FACIL exhibit the highest contributions, underscoring their relevance in the classification model. The error bars in the image denote the variability in importance across multiple runs.

4.2 Assessing Model performances

To address the research questions we previously posed, we need to quantitatively assess how accurately the random forest model can solve the binary classification task defined in Section 3.5. To this end, we need a standard measure to assess how effectively a port is highly central in the graph of vessel voyages by examining its features alone **(RQ1)**. Moreover, to interpret the importance of features in solving this ML task, we need to ensure that the model identifies a pattern in the data instead of returning random guesses.

The area under the *Receiver Operating Characteristic* (ROC) curve is a widely used indicator of prediction accuracy for binary classification problems. The ROC curve is obtained by exploring various thresholds for a model's binary classification and plotting the True Positive rate versus the False Positive rate. The resulting curve describes the performance of the binary classifier at hand, and the area under it is referred to as the Area Under the Curve (AUC). In short, the closer the ROC is to the top left angle, the better the

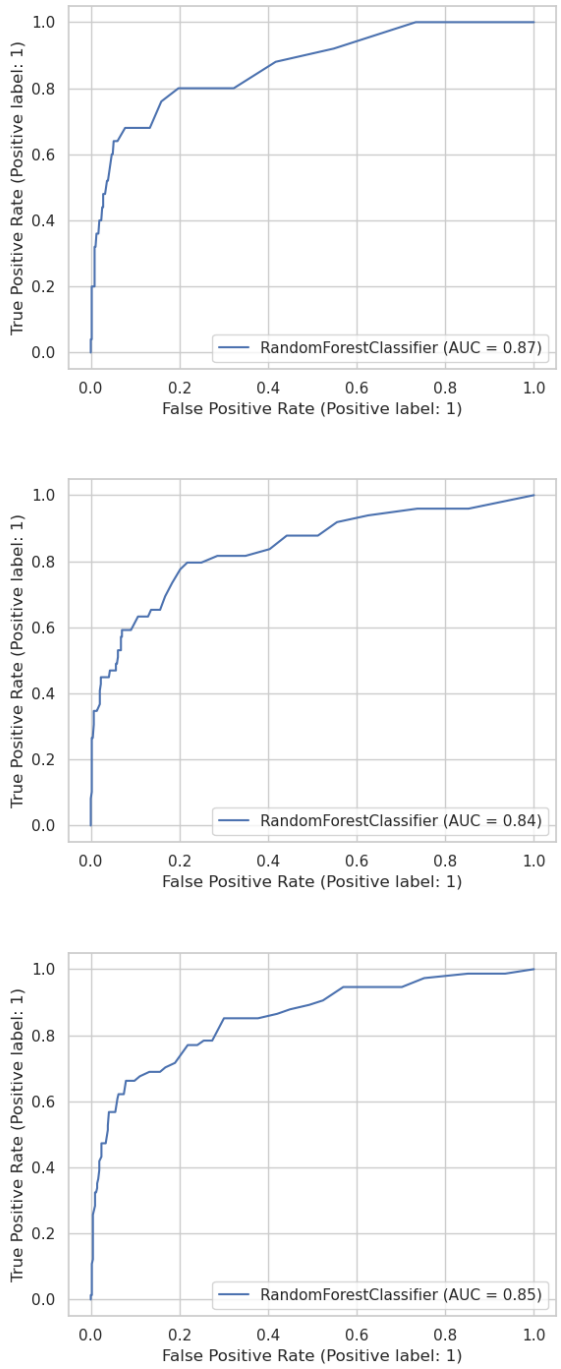

Figure 4: ROC curve from the random forest for the binary classification of most central ports. From top to bottom, we consider relevant ports, the ones on the 5%**,** 10%**,** 15%

predictions, *i.e.*, AUC close to 1. For reference, a random classifier would be close to the diagonal line, with an AUC of 0.5.

For all three threshold configurations for the most central ports (5%, 10%, 15%), we show in Figure 4 the ROC curve for the binary classification of the most central ports, and the AUC in Table 4. The results confirm that the random forests obtain good results in the classification task and that the following analysis of feature importance is based on an accurate classifier, as the values of the AUC obtained are well above the 0.5 (0.87, 0.84, 0.85, respectively) value that would be received by random classification.

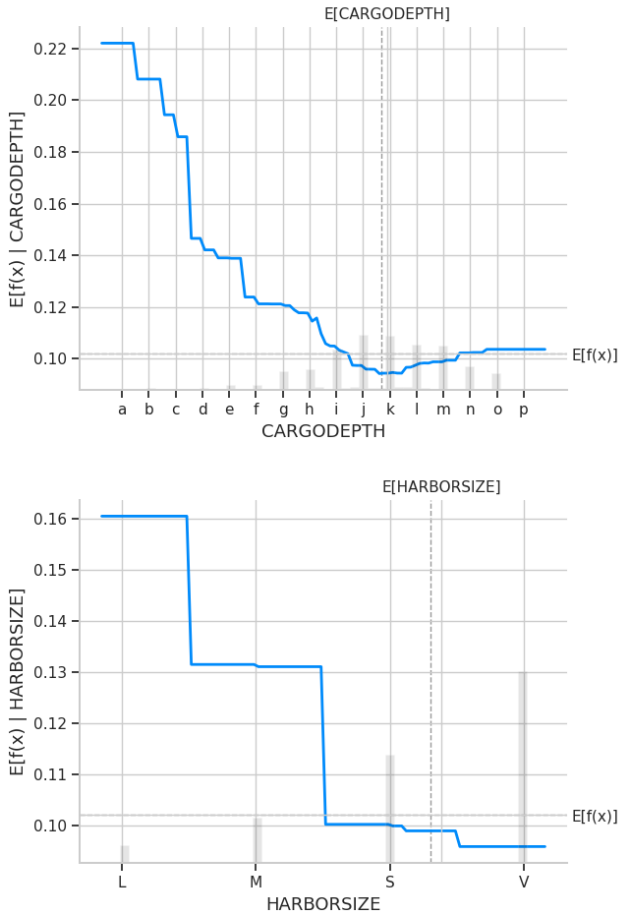

Figure 5: Partial dependence plots for the most important features according to SAGE, CARGODEPTH (top) and HARBOURSIZE (bottom). The vertical gray bar represents the average value of the feature. The blue partial dependence plot line is the average value of the model output when we fix the feature at hand to a given value. The grey histograms on the x-axis indicate the distribution of each feature.

4.3 Feature Importance

Subsequently, we analyze the results of the feature importance from the random forest model **(RQ2)**, discussing predictability and co-dependence between the aggregated centrality and port features.

4.3.1 Global feature importance. In Figure 3, we show the SAGE values describing feature importance for the whole Port Network when we considered the top 10% of the ports as central. It becomes

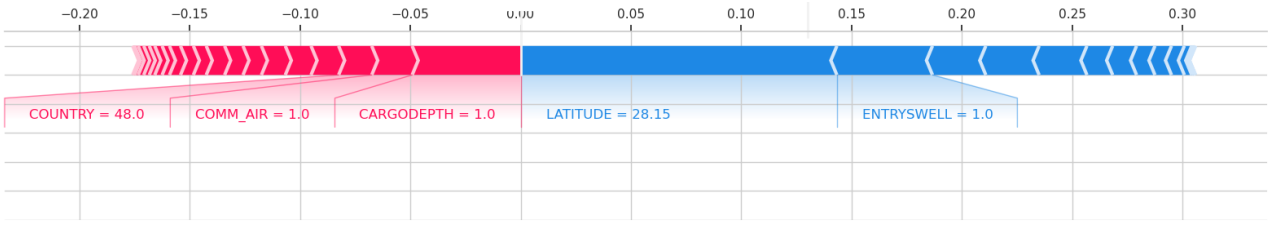

Figure 6: SHAP Force Plot for Feature Contributions at Las Palmas Port. This force plot visualizes the impact of various features on the classification outcome for the Las Palmas port case. Features contributing to a correct classification are highlighted in red, while those contributing to an incorrect classification are shown in blue. Notable features include COUNTRY, COMM_AIR, CARGODEPTH, LATITUDE, and ENTRYSWELL, with their respective values influencing the model's prediction.

clear that the most important features in predicting port centrality are CARGODEPTH, LONGITUDE, and HARBORSIZE.

CARGODEPTH indicates the maximum depth available for cargo vessels in the port. A greater depth allows for large vessels to visit the port. For example, cargo vessels require a depth greater than 12 meters, up to 25 meters for deeper ships. The dataset's depth is coded using letters from A to Q, with Q representing the deepest level. Each subsequent letter increases the depth by 5 feet, approximately 1.5 meters. For example, the letter H corresponds to 43 feet (≈13 meters). The information about the depth is the most useful for discriminating whether a cargo ship can enter the port.

The HARBORSIZE is based on several factors, including area, facilities, and wharf space. It is codified into four categories: Large (L), Medium (M), Small (S), or Very Small (V). Similar to cargo depth, most cargo visits are focused on large ports, which makes this feature one of the most informative. Finally, LONGITUDE shows the importance of a port's geographical position. It is interesting to note how longitude is more relevant than latitude in this context. One possible explanation is that most central ports, as graphically illustrated in Figure 2, are located in specific longitudinal areas, including the Americas, Europe, and Southeast Asia. Longitude increases accessibility and is a good discriminator for central ports.

While SAGE quantifies the importance of a feature in improving the model's prediction, it does not provide information on the effect of different values of each variable on the model's prediction. To gain additional insight, Figure 5 shows the partial dependency plots for the features that were globally most important. These plots illustrate how the model's prediction changes on average when the value of one variable is altered. Notice that the dependencies are highly non-linear due to the non-linear nature of the model we employed. Moreover, the CARGODEPTH data means higher values correspond to lower values in meters. Hence, increasing the actual depth increases the likelihood that a port will be predicted as highly central. Finally, the correlation between a harbor size (HARBORSIZE) and port centrality is made evident in the plot.

Most Central Ports Threshold	5%	10%	15%
AUC value	0.87	0.84	0.85
CARGODEPTH	0.02	0.04	0.08
LONGITUDE	0.07	0.04	0.07
HARBORSIZE	0.03	0.02	0.04
COUNTRY	0.01	0.03	0.01
HAN_DEPTH	0.02	0.02	0.006

Table 4: The 5 most important features and their SAGE values for predicting the most central ports in a binary classification

4.3.2 Local Feature Importance. Local feature importance measures the relevance of a feature in classifying single-input items. To summarize what features are more locally important, we ranked them according to their average contribution to each port's classification. Intuitively, the average ranking would measure how consistently a feature contributes to a port's positive classification. In general, the geographical features ranked the highest. Apart from those, Table 5 shows the top 5 non-geographical features. Interestingly, a "political" feature ranks first. Then, services such as medical procedures and management of contaminated materials are typical of international ports and could differentiate the port centrality. Finally, port depths directly correlate with the port's ability to accommodate large cargo ships, a further centrality indicator.

Another interesting analysis involves examining individual ports and how their features correlate with their local importance. Figure 6 shows the Las Palmas port. CARGODEPTH, COMM_AIR (indicates whether airport communications are available), and COUNTRY are features that help classify them. Contrarily, its latitude and ENTRY SWELL (binary, whether there is a natural factor restricting the entrance of vessels) are not favorable features for this port.

5 Conclusion

We use explainable machine learning to identify key features of significant ports. To achieve this, we constructed a Ports Network using three years of worldwide data, where the significance of a port is determined by the combination of centrality measures commonly used in the literature. We performed a machine learning task to predict the port's importance using publicly available features and analyzed which features are most useful for inference.

Geographical features are the most informative for identifying central ports, *i.e.*, a direct and expected outcome, given that ports are inherently tied to specific coastal locations. Beyond geography, features related to port depth, such as entrance and pier depth, also

Feature	Avg. Rank	Description
US REPRESENTATIVE	9.8	Indicates whether the United States maintains either civilian or military representation in the port.
DIRTY BALLAST	12.2	Whether a port has sufficient facilities for receiving oily or contaminated ballast.
PRATIQUE	12.6	Whether medical practice is applied to vessels arriving in the port.
CARGODEPTH	13.8	The Greatest depth for cargo vessels available in the port.
CHANNEL DEPTH	14.3	The depth of the deepest channel leading to the port.

Table 5: Top 5 non-geographical features ordered by their average rank

play a critical role in accurate classification. This aligns with expectations, as the analysis focuses on cargo vessels, which typically require deep-water infrastructure for safe access and docking.

This paper also paves the way for similar research on other types of vessel modalities, such as passenger or leisure vessels. Additionally, identifying the features of central maritime ports can facilitate research on developing new or existing ports and provide a tool for studying inter-port and regional relationships. These points provide potential research directions for extending this study.

Acknowledgments

The authors acknowledge the support of the H2020 EU Project MASTER (Multiple ASpects TrajEctoRy management and analysis), funded under the Marie Skłodowska-Curie grant agreement No. 777695. This research was partially supported by the Institute for Big Data Analytics (IBDA) and the Ocean Frontier Institute (OFI) at Dalhousie University, Halifax, NS, Canada, as well as by the Canadian Foundation for Innovation's MERIDIAN Cyberinfrastructure. Additional funding was provided by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), Brazil.

Data Licensing and Disclosure

The research described in this paper used data acquired from Spire under MERIDIAN's fair-use and non-disclosure data license. Due to licensing agreements, we are unable to share the raw data. However, we can provide a trained model and the means to retrain the models using open-source data from similar maritime regions.

References

[1] M. Alam, Gabriel Spadon, Mohammad Etemad, Luis Torgo, and E. Milios. 2024. Enhancing short-term vessel trajectory prediction with clustering for heterogeneous and multi-modal movement patterns. *Ocean Engineering* 308 (Sept. 2024), 118303. doi:10.1016/j.oceaneng.2024.118303

[2] N. G. Álvarez, N. Adenso-Díaz, and L. Calzada-Infante. 2021. Maritime traffic as a complex network: A systematic review. *Netw. and Spatial Econ.* (2021), 1–31.

[3] Emanuele Carlini, Vinicius Monteiro de Lira, Amilcar Soares, Mohammad Etemad, Bruno Brandoli, and Stan Matwin. 2021. Understanding evolution of maritime networks from automatic identification system data. *GeoInformatica* (2021), 1–25.

[4] E. Carlini, V. Monteiro de Lira, A. Soares, M. Etemad, B. Brandoli Machado, and S. Matwin. 2020. Uncovering vessel movement patterns from AIS data with graph evolution analysis. In *EDBT/ICDT Workshops*.

[5] Kam-Fung Cheung, Michael GH Bell, Jing-Jing Pan, and Supun Perera. 2020. An eigenvector centrality analysis of world container shipping network connectivity. *Trans. Research Part E: Log. and Trans. Rev.* 140 (2020), 101991.

[6] Ian Covert, Scott M Lundberg, and Su-In Lee. 2020. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems* 33 (2020), 17212–17223.

[7] G. Del Mondo, P. Peng, J. Gensel, C. Claramunt, and F. Lu. 2021. Leveraging Spatio-Temporal Graphs and Knowledge Graphs: Perspectives in the Field of Maritime Transportation. *ISPRS Int. J. of Geo-Information* 10, 8 (2021), 541.

[8] C. Ducruet, S. Lee, and A. KY Ng. 2010. Centrality and vulnerability in liner shipping networks: revisiting the Northeast Asian port hierarchy. *Maritime Policy & Management* 37, 1 (2010), 17–36.

[9] C. Ducruet, C. Rozenblat, and F. Zaidi. 2010. Ports in multi-level maritime networks: evidence from the Atlantic (1996–2006). *J. of Trans. Geo.* 18, 4 (2010), 508–518.

[10] Zuzanna Kosowska-Stamirowska, César Ducruet, and Nishant Rai. 2016. Evolving structure of the maritime trade network: evidence from the Lloyd's Shipping Index (1890–2000). *Journal of Shipping and Trade* 1, 1 (2016), 1–17.

[11] F. G. Laxe, M. J. F. Seoane, and C. P. Montes. 2012. Maritime degree, centrality and vulnerability: port hierarchies and emerging areas in containerized transport (2008–2010). *J. of Trans. Geo.* 24 (2012), 33–44.

[12] J. Li, X. Wang, and T. Zhang. 2021. Sequence-based centrality measures in maritime transportation networks. *IET Int. Trans. Sys.* 14, 14 (2021), 2042–2051.

[13] Y. Liu and S. D. Brown. 2013. Comparison of five iterative imputation methods for multivariate classification. *Chemometrics and Int. Lab. Sys.* 120 (2013), 106–115.

[14] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[15] L. H. McWhinnie, P. D. O'Hara, C. Hilliard, N. Le Baron, L. Smallshaw, R. Pelot, and R. Canessa. 2021. Assessing vessel traffic in the Salish Sea using satellite AIS: An important contribution for planning, management and conservation in southern resident killer whale critical habitat. *Ocean & Coastal Management* 200 (2021), 105479. doi:10.1016/j.ocecoaman.2020.105479

[16] C. P. Montes, M. J. F. Seoane, and F. G. Laxe. 2012. General cargo and containership emergent routes: A complex networks description. *Trans. Pol.* 24 (2012), 126–140.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[18] F. A. Rodrigues. 2019. Network centrality: an introduction. In *A mathematical modeling approach from nonlinear dynamics to complex systems*. Spr., 177–196.

[19] M. J. F. Seoane, Fernando G. Laxe, and C. P. Montes. 2013. Foreland determination for containership and general cargo ports in Europe (2007–2011). *J. of Trans. Geography* 30 (2013), 56–67.

[20] Lloyd S Shapley. 1953. A value for n-person games, Contributions to the Theory of Games, 2, 307–317.

[21] R. Song, G. Spadon, R. Pelot, S. Matwin, and A. Soares. 2024. Enhancing global maritime traffic network forecasting with gravity-inspired deep learning models. *Scientific Reports* 14, 1 (2024), 16665. doi:10.1038/s41598-024-67552-2

[22] G. Spadon, J. Kumar, D. Eden, J. van Berkel, T. Foster, A. Soares, R. Fablet, S. Matwin, and R. Pelot. 2024. Multi-path long-term vessel trajectories forecasting with probabilistic feature fusion for problem shifting. *Ocean Engineering* 312 (2024), 119138. doi:10.1016/j.oceaneng.2024.119138

[23] M. Sundararajan, A. Taly, and Q. Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*. PMLR, 3319–3328.

[24] D. Tocchi, C. Sys, A. Papola, F. Tinessa, F. Simonelli, and V. Marzano. 2022. Hypergraph-based centrality metrics for maritime container service networks: A worldwide application. *J. of Trans. Geo.* 98 (2022), 103225.

[25] B. Tovar, R. Hernández, and H. Rodríguez-Déniz. 2015. Container port competitiveness and connectivity: The Canary Islands main ports case. *Trans. Pol.* 38 (2015), 40–51.

[26] Nguyen Khoi Tran and Hans-Dietrich Haasis. 2014. Empirical analysis of the container liner shipping network on the East-West corridor (1995–2011). *NETNOMICS: Economic Research and Electronic Networking* 15, 3 (2014), 121–153.

[27] Iraklis Varlamis, Ioannis Kontopoulos, Konstantinos Tserpes, Mohammad Etemad, Amilcar Soares, and Stan Matwin. 2021. Building navigation networks from multi-vessel trajectory data. *GeoInformatica* 25, 1 (2021), 69–97.

[28] Iraklis Varlamis, Konstantinos Tserpes, Mohammad Etemad, Amílcar Soares Júnior, and Stan Matwin. 2019. A Network Abstraction of Multi-vessel Trajectory Data for Detecting Anomalies.. In *EDBT/ICDT Workshops*.

[29] Y. Wang and K. Cullinane. 2016. Determinants of port centrality in maritime container transportation. *Trans. Research Part E: Log. and Trans. Review* 95 (2016), 326–340.

[30] Zhihuan Wang, Christophe Claramunt, and Yinhai Wang. 2019. Extracting global shipping networks from massive historical automatic identification system sensor data: a bottom-up approach. *Sensors* 19, 15 (2019), 3363.

[31] Dong Yang, Lingxiao Wu, Shuaian Wang, Haiying Jia, and Kevin X Li. 2019. How big data enriches maritime research–a critical review of Automatic Identification System (AIS) data applications. *Transport Reviews* 39, 6 (2019), 755–773.

[32] Fan Zhang, Yihao Liu, Lei Du, Floris Goerlandt, Zhongyi Sui, and Yuanqiao Wen. 2023. A rule-based maritime traffic situation complex network approach for enhancing situation awareness of vessel traffic service operators. *Ocean Engineering* 284 (2023), 115203. doi:10.1016/j.oceaneng.2023.115203