12º Seminário Internacional de Transporte e Desenvolvimento Hidroviário Interior

Rio de Janeiro/RJ, 19th to 21st October 2021

Ocean Port Congestion Indicators - A Machine Learning Approach

Vinícius Barreto Martins, UFRJ/COPPE, Rio de Janeiro/Brasil, v1nybarreto@oceanica.ufrj.br
Ramiro Fernandes Ramos, UFRJ/COPPE, Rio de Janeiro/Brasil, ramirofr@oceanica.ufrj.br
Maricruz Aurelia Fun Sang Cepeda, UFRJ/COPPE, Rio de Janeiro/Brasil, maricruzcepeda@oceanica.ufrj.br
Jean-David Caprace, UFRJ/COPPE, Rio de Janeiro/Brasil, jdcaprace@oceanica.ufrj.br

Abstract

Maritime trade plays a crucial role in the global economy, and recent technological developments have accelerated marine logistics. However, this increase impacted port performance, leading to port congestion in some regions and distorting the smooth flow of maritime logistics. Few studies employing AIS data have explored marine traffic congestion; hence, developing a system that makes port metrics more accessible is needed. This work employs a methodology to analyze the port congestion level of Rio de Janeiro. Three algorithms were developed using the Automatic Identification System (AIS) data to identify the geolocation area, the convex hull area, and the average vessel's proximity. These algorithms were used to calculate the Port Congestion Indicators (PCIs): spatial concentration, spatial density, average service time, and Machine Learning techniques were employed to extract knowledge from the database. As a result, this process identified the periods when ports are most congested, and the centroids of these clusters can be used to predict future congestion levels. These indicators provide resources for better management and can motivate actions such as the redistribution of ship loading and unloading locations and improving port performance measurement.

1. Introduction

It is estimated that maritime trade represents 90% of the global volume trade and, therefore, the port's performance is crucial to sustaining economic growth [1]. However, this increase in maritime trade has had an impact on the port's efficiency in some regions. In December 2019, for example, ships that operated liquid chemical bulk in the Port of Santos had to wait more than ten days for a docking opportunity, causing losses of around $US\$\ 35.000$ per day for each ship [2].

This situation is called port congestion, in which vessels have to wait at anchorage areas before accessing the port for load or unload. This impact is not restricted to any part of the world affecting ports in Asia, North Africa, Northern Europe, and the United States [3]. Port congestion is an important issue from an economic and efficiency point of view. Because it results, not only, in longer waiting times and low service levels for vessels, but it also contributes to the decrease in competitiveness and demand [3].

Understanding the aspects that influence congestion is essential for port management. However, traditional traffic analyzes are, generally, carried out through surveys that include: visual observations; radar, and aerial photographs, being extremely costly [4]. Currently, there are advanced methods for collecting vessel traffic data, such as

Vessel Traffic Services (VTS) and Automatic Identification System (AIS).

The Automatic Identification System (AIS) is a vessel tracking system that provides regular ship data updates. Static information and dynamic information of the vessels can be exchanged electronically between AIS receiving stations [1]. AIS data does not have the limitations of VTS data and, due to its informative integrity, can be used to analyze incidents, such as ship collisions.

This data ensures greater reliability of navigation information and comes to the analysis of maritime traffic more accurately [4] [5]. However, most studies performed with AIS data have focused on specific areas such as monitoring, tracking, security of ships, accident prevention, including collision risks, noise levels, or ship emissions [6].

Few kinds of research employing AIS data have explored marine traffic congestion. Inspired by the work of Craighead et al. (2007) [7], AbuAlhaol et al. (2018) [1] proposed three "Big Data-Driven" indicators to measure the marine traffic congestion: the spatial density of seaports; the spatial complexity; and the average waiting time for ships. From these indicators, the K-means clustering technique was used to identify the periods in which the selected ports were more or less congested, motivating actions such as the redistribution of ships loading and unloading locations and better anchorage planning [1].

This work complements the methodology proposed by AbuAlhaol et al. (2018) [1] to analyze the port congestion level of Rio de Janeiro port. Over one year, AIS data were collected and analyzed to calculate the Port Congestion Indicators (PCIs): spatial concentration, spatial density, average service time. Machine learning techniques were applied to identify the periods in which the port is most congested, allowing predicting the future status.

This model could increase the port's efficiency and be applied to other regions, providing resources to port authorities for better management, improving the port logistics. Also, this work is inspired by Martins (2021) [8], the master's thesis presented to Ocean Engineering Program at COPPE/UFRJ and Martins (2020) [9], an article published at the 28th International Congress on Waterway Transport, Shipbuilding, and Offshore.

2. AIS Database

A database is a collection of information - preferably related information and organized. A database is a structured object which consists of data and metadata. Data in a database is the actual stored descriptive information, and metadata describes the structure applied to the database [10]. The database used in this work is the AIS messages transmitted by ships in the region of the port of Rio de Janeiro. The data consists of January 2018 to April 2018 and September 2018 to March 2019, containing 141 million records and 196 attributes.

Inaugurated on 20th July 1910, the port of Rio de Janeiro, located on the west coast of Guanabara Bay, in the city of Rio de Janeiro (latitude: $-22° 53' 31'' S$, longitude: $-43° 11' 43'' W$), works with potential cargoes such as general containerized cargo, electronics, rubber, petrochemicals, vehicle parts, coffee, steel products, press paper reels, and solid bulk such as wheat and pig iron. In 2016, the port managed a total of $6.102.907\ tons$ and $299.833\ TEU$ of cargo and container, respectively [11].

The AIS data transmitted on the region of the port of Rio de Janeiro were structured in the SQL Server (relational database management system). Of the 196 attributes, just twelve were selected to hasten the queries: Message-ID; User ID; Navigational Status; Speed Over Ground; Longitude; Latitude; IMO Number; Name; Type of Ship; Vessel Length; Vessel Beam; and Date Time Stamp. To become the analysis more accurate, the periods in which AIS data were not transmitted or may have been some technical issues were removed (see Table 1). The next section describes the methodology used in this work.

Table 1: Data removed from the Rio de Janeiro port.

Port	ID	Period (DD/MM/YY)	Records
Rio de Janeiro	5	$09/02/2018 - 16/02/2018$	0
	15	$25/04/2018 - 31/04/2018$	0
	31	$25/12/2018 - 31/12/2018$	0
	32	$01/01/2019 - 08/01/2019$	0
	43	$25/03/2019 - 31/03/2019$	0

3. Methodology

This section describes the methodology used in this work. The first step is to collect and prepare the AIS data transmitted by ships in the region. The next step is to calculate the geospatial algorithms: convex hull area, geolocation area, and average vessel proximity. Then, these algorithms will be used to calculate the normalized Port Congestion Indicators (PCIs): spatial concentration, spatial density, and average service time. Finally, unsupervised and supervised learning algorithms will be employed to extract useful information.

3.1 Data Processing

Data processing refers to transforming raw data into meaningful output i.e., information [12]. The data processing cycle refers to the sequence of activities involved in data transformation, and this process can be divided into six stages: collection, preparation, input, processing, output, and storage [13] [14]. The collection is the first stage of the cycle and is crucial since the quality of data collected will heavily impact the output [12] [14].

The next step is the preparation of the data into a form suitable for further analysis and processing [14]. Preparation is about constructing a data set from one or more data sources to be used for further exploration and processing. Input is the task where verified data is converted into a machine-readable form so that it can be processed through an application.

Processing is when the data is subjected to various means and methods of powerful technical manipulations using Machine Learning and Artificial Intelligence algorithms to generate an output or interpretation. Output and interpretation are the stages where processed information is transmitted and displayed to the user [14].

Storage is the last stage in the data processing cycle, where data and metadata (data information) are held for future use. The importance of this cycle is that it allows quick access and retrieval of the processed information, allowing it to be passed on to the next stage directly when needed. Also, adequately stored data is a necessity for compliance [14].

3.2 Relational Database

Values represent all information in a relational database in tables (even table names appear as character strings in at least one table). Addressing data by value, rather than by position, boosts the productivity of programmers and end-users (positions of items in sequences are usually subject to change and are not easy for a person to keep track of, especially if the sequences contain many items. The relational database is best suited to data with a somewhat stable or homogeneous structure [15].

The relational data model is the result of the work of Edgar Codd. During the 1960s, Dr. Codd, although trained as a mathematician, worked with existing data models. His experience led him to believe that the ways of representing data relationships were clumsy and unnatural. Therefore, he went back to mathematical set theory and focused on the construct known as a relation. He extended that concept to produce the relational database model, which he introduced in a paper in 1970 [16].

In mathematical set theory, a relation is the definition of a table with columns (attributes) and rows (tuples). The report specifies what will be contained in each column of the table but does not include data. When you include rows of data, you have an instance of a relation [16].

The two main relational languages are SQL and QBE, with SQL being the most important [17]. Structured Query Language, or SQL, is a programming language used in the relational database management system, developed by IBM in the early 1970s [18] [19]. A relational database management system (RDBMS) is a program that allows you to create, update, and administer a relational database using the SQL language. The most popular RDBMS are SQL Server, MySQL, PostgreSQL, Oracle DB, and SQLite [20].

Microsoft SQL Server is a relational database management system that supports various transaction processing, business intelligence, and analytics applications. In our work, the SQL server was used to manages the AIS data gathered over the period, containing attributes and records. To fasten the queries, the essential attributes for the calculation of the geospatial algorithms and port congestion indicators were selected, which will be presented in the following sections.

3.3. Geospatial Algorithms

The term geospatial defines the collective data and associated technology that has a geographic or locational component. It means that the records in a dataset have locational information attached to them, such as geographic data containing coordinates [21]. The geospatial technologies describe the range of modern tools contributing to geographic mapping and analysis [22]. For processing the geographic data, the technologies used in this work are the geospatial algorithms for the convex hull area, geolocation area, and average vessel proximity.

A. Convex Hull Area

The convex hull of a set of points is the smallest convex set that contains the points, it is a fundamental construction for mathematics and computational geometry [23]. Quickhull is an algorithm for computing convex hulls that takes a divide-and-conquer approach. The idea is to partition the problem into subproblems of roughly equal size, solve each subproblem recursively, and finally combine the individual results into a whole solution [24]. The port convex hull area is defined as enclosing all vessels in the smallest perimeter fence [1].

B. Geolocation Area

Geohash is an address code that reduces two-dimensional latitude and longitude information to a unique one-dimensional string for each point on earth with a given precision [25]. In our work, we split the convex area into cells of a precision factor of seven, and the coordinates of these cells were used to identify which area is been used by the vessels. Each cell can only be activated once in the period if there is at least one vessel within the coordinates of that region. Then, to calculate the geolocation area, we considered the area of each activated cell over the period.

C. Average Vessels Proximity

The average vessel proximity is the average distance between the location of all vessels [26]. The input to the algorithm is the latitude and longitude of the ships. The first step is a combination, without repetition, of the vessel's coordinates. Then, the algorithm calculates the distance between the coordinate pairs. The interquartile range (IQR) method was applied to remove the outliers that account for the effect of immense distances.

3.4. Port Congestion Indicators

This section provides a detailed formulation of three Port Congestion Indicators (PCIs) that capture the spatial concentration, spatial density, and average service time of the period and ports of interest, based on the geospatial algorithms. The term congestion emphasizes the fact that high indicator values magnify the significance of port disruption. Therefore, the port with high PCIs is more vulnerable [1].

A. Spatial Concentration

Spatial Concentration (SC) is the normalized port congestion indicator to measure the spatial distribution of ships within the convex area. It is calculated by dividing the Convex Hull Area ($Convex\ Area_{(i)}$) by the Average Vessels Proximity ($\Delta_{(i)}$).

Equation 1: Spatial Concentration Indicator.

$$SC_{(i)} = \frac{Convex\ Area_{(i)}/\Delta_{(i)}}{\max\limits_{i \in I} \{Convex\ Area_{(i)}/\Delta_{(i)}\}}$$

B. Spatial Density

Spatial Density (SD) is the normalized port congestion indicator to measure the area used by the vessels. It is calculated by dividing the Geolocation Area ($Geo\ Area_{(i)}$) by the Convex Hull Area ($Convex\ Area_{(i)}$).

Equation 2: Spatial Density Indicator.

$$SD_{(i)} = \frac{Geo\ Area_{(i)}/Convex\ Area_{(i)}}{\max\limits_{i \in I} \{Geo\ Area_{(i)}/Convex\ Area_{(i)}\}}$$

C. Average Vessels Proximity

The third Port Congestion Indicator (i.e., Average Service Time) represents the average time needed by vessels to enter, load/offload, and exit the port over the period. We define t_n as the time needed for a vessel to get served and leave the port, N_i the number of unique vessels (based on the reported MMSI numbers) in the i_{th} aggregation period.

Equation 3: Average Service Time Indicator.

$$AST_{(i)} = \frac{\frac{1}{N_i}\ x\ \sum_{n=1}^{N_i} t_n}{\max\limits_{i \in I} \left\{\frac{1}{N_i}\ x\ \sum_{n=1}^{N_i} t_n\right\}}$$

3.5. Machine Learning

A key characteristic of machine learning is the concept of self-learning. This refers to applying statistical modeling to detect patterns and improve performance based on data and practical information [27]. The application of machine learning methods to large databases is called data mining, in which a large volume of data is processed to construct a simple model with practical use. However, machine learning is not just a database problem; it is also a part of artificial intelligence [28]. There are so many different applications of machine learning that it is helpful to classify them. The first category, and employed in this work, is based on whether or not they are trained with human supervision (unsupervised or supervised learning algorithms) [29].

A. Unsupervised Learning Algorithms

In unsupervised learning, the machine simply receives inputs $x1$, $x2$, but obtains neither supervised target outputs nor rewards from its environment. However, it is possible to develop a formal framework for unsupervised learning based on the notion that the machine's goal is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine. In a sense, unsupervised learning can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise [30].

For clustering, we need to identify distinct groups such that the instances within a group are similar to each other but different from instances in other groups. One such algorithm is K-means clustering. An even more powerful clustering algorithm, based on the density of points, is known as DBSCAN (density-based spatial clustering of applications with noise). The algorithm will group those that are packed closely together [31]. These algorithms will be discussed and employed in Section 4.1, for the identification of the congested periods.

B. Supervised Algorithms

Supervised machine learning involves predetermined output attributes besides the use of input attributes. The algorithms attempt to predict and classify the predetermined attribute. Their accuracies and misclassification alongside other performance measures depend on the counts of the predetermined attribute correctly predicted or classified or otherwise. It is also important to note the learning process stops when the algorithm achieves an acceptable level of performance [32].

Supervised algorithms perform analytical tasks first using the training data and subsequently construct contingent functions for mapping new instances of the attribute. The algorithms require pre-specifications of maximum settings for the desired outcome and performance levels. Given the approach used in machine learning, it has been observed that a training subset of about 66% is a rationale and helps in achieving the desired result without demanding more computational time [32].

Supervised learning is further subdivided into classification and regression. For cases where we only have a few predefined labels to predict, we use a classifier. However, if we want to predict a wide-range number, it is a regression problem since these values can be anything [29].

Classification can be performed on structured or unstructured data and is a technique to categorize data into a given number of classes. The main objective is to identify the category or class to which new data will comprise. There are several types of classification algorithms, depending on the dataset. The most common supervised learning algorithms are K-Nearest Neighbors, Random Forest, Logistic Regression, Naive Bayes Classifier, and Support Vector Machines [33] [34].

The K-Nearest Neighbor classification is one of the most fundamental methods when it is little or no prior knowledge about the distribution of the data [35]. Random forests are a combination of tree predictors. Each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [36]. These algorithms will be discussed and employed in section 4.2, for the prediction of the congestion levels.

4. Results and Discussions

This section presents the results and discussions concerning unsupervised and supervised learning algorithms for knowledge discovery in databases. In section 4.1 were used the K-means and DBSCAN algorithms were to cluster the normalized Port Congestion Indicators (PCIs) and identify the congested periods. In section 4.2 were used the K-Nearest Neighbors and Random Forest algorithms to predict the congestion level of Rio de Janeiro port.

4.1. Clustering

This section details the employment of unsupervised learning algorithms (K-means and DBSCAN) to cluster the normalized Port Congestion Indicators (PCIs) spatial concentration, spatial density, and average service time, previously calculated to identify the congested periods. The K-means algorithm was performed, setting $K = 3$ to aggregates the periods into three clusters. The cluster centroid closest to $(1, 1, 1)$ represents the highest congestion. (i.e., close to the maximum port congestion indicators values). The cluster centroid most distant to $(1, 1, 1)$ represents the lowest congestion. Finally, the cluster centroid between the closest and most distant to $(1, 1, 1)$ illustrates the medium congestion.

The DBSCAN algorithm was performed, setting the $MinPts$ ($K = 2$). Once the $MinPts$ value is set, were calculated the Best Eps Value (the maximum point of curvature or greatest slope on a graph) for each scenario. Since the algorithm is designed to cluster data of arbitrary shapes in the presence of noise, the representation of the indicators closest to $(1, 1, 1)$ in decrescent order is high, medium-high, medium, medium-low, low, and noise if do not belong to any cluster. Next, Table 2 presents the PCIs clustering for Rio de Janeiro port, which resumes the periods that were considered as high congestion, for both algorithms (K-means and DBSCAN), which could require some actions.

Table 2: Identification of the Congested Periods.

Port	ID	Period (DD/MM/YY)
Rio de Janeiro	2	$17/01/2018 - 24/01/2018$
	9	$09/03/2018 - 16/03/2018$
	10	$17/03/2018 - 24/03/2018$
	37	$09/02/2019 - 16/02/2019$
	6	$01/10/2018 - 31/10/2018$
	10	$01/02/2019 - 31/02/2019$

The port congestion indicators could motivate the authorities to take special consideration at the high congested periods. Abualhaol et al. [1] recommend redistributing the anchored vessels, expanding and developing the port's infrastructure to deal with increasing demand. Also, could allocate more resources to speed up the loading and unloading of ships, considering a redesign of operations and internal processes.

Despite the complexity, some attempts have been made to try and ease the congestion. A free-flow study conducted at the Port of Los Angeles in early 2015 found that free-flowed containers reduced trucker turn-times by more than 50 percent — from 85 minutes to 42 minutes, eliminating unnecessary movements. This system lets the ports peel off the first available container and place it on the next available driver [37].

In 2018 Ports in Oakland introduced night shifts. It compensates for the delay and expedites the clearing of ships from the ports. An alternative solution would be using SOC Containers instead of COC Containers because they don't need to be returned to the port but instead to the owner. This might not eliminate port congestion, but it helps avoid demurrage and detention charges for the cargo owner [38].

Another strategy for high congestion periods is cargo redistribution. Companies can achieve better supply balance by negotiating with various ocean carriers with terminal operators in multiple ports. This can provide viable options for inbound and outbound supply chains without significantly impacting costs. Also, strategically managing the virtual warehouse by monitoring supply and demand during inland transit at the origin or destination could yield greater flexibility [39].

Port congestion is an increasingly dangerous threat to maritime logistics. It forces the companies to increase operational costs, losing their credibility to transport on time. Also, the cargo can miss their connecting ships or trucks to different destinations, causing scheduling problems. Therefore, it is essential to monitor the port performance continually and mitigate the effects at the high congested periods. The following section employs supervised learning algorithms for predicting the congestion level at Rio de Janeiro port.

4.2. Prediction

This section details the employment of supervised learning algorithms (K-Nearest Neighbors and Random Forest) to predict the congestion level of Rio de Janeiro port, based on the previous classification of the K-means and DBSCAN algorithms. For forecasting, we have some labeled data that is used for training the model and unlabeled data that we want to classify.

The first step is to separate the training subset from the test subset. Most of the algorithms require that the attributes and the labels are in separate variables. Given the approach used in machine learning, were employed a training subset about $70\% - 80\%$ and a test subset about $20\% - 30\%$. The x training subset contains the normalized attributes spatial concentration, spatial density, and average service time. The y training subset contains the labeled data classified by K-means and DBSCAN algorithms.

For the K-Nearest Neighbors algorithm, were select the best value of K and made predictions on x test subset to predict the congestion level. The optimal choice of the K value is highly data-dependent, and we used $K = 2$, given the little neighborhoods. For the Random Forest algorithm, were used the default values of the algorithm. These parameters, controlling the size of the trees, lead to fully grown and unpruned trees, which can potentially be very large on some data sets, increasing the memory consumption, the complexity, and the size of the trees.

The principle behind the K-Nearest Neighbor methods is to find a predefined number of training samples closest in the distance to the new point and predict the label from these. The number of samples can be a user-defined constant or vary based on the local density of points. Neighbors-based methods are known as non-generalizing machine learning methods since they simply "remember" all of their training data [40].

Random forests are considered a highly accurate and robust method because of the number of decision trees participating in the process. It takes the average of all the predictions, which cancels out the biases [41]. A significant disadvantage of random forests lies in their complexity. Next, Table 3 and Table 4, presents the PCIs predicting for Rio de Janeiro port by week, based on the previous classification of the K-means and DBSCAN algorithms. The accuracy of the supervised learning algorithms is represented by "green" for correct prediction and "red" for incorrect.

Table 3: Prediction of the congestion level (K-means).

ID	Period (DD/MM/YY)	KNN	R. Forest
33	09/01/2019 – 16/01/2019	LOW	LOW
34	17/01/2019 – 24/01/2019	HIGH	HIGH
35	25/01/2019 – 31/01/2019	MEDIUM	MEDIUM
36	01/02/2019 – 08/02/2019	HIGH	HIGH
37	09/02/2019 – 16/02/2019	HIGH	HIGH
38	17/02/2019 – 24/02/2019	HIGH	HIGH
39	25/02/2019 – 31/02/2019	MEDIUM	MEDIUM
40	01/03/2019 – 08/03/2019	MEDIUM	MEDIUM
41	09/03/2019 – 16/03/2019	MEDIUM	MEDIUM
42	17/03/2019 – 24/03/2019	LOW	LOW

Table 4: Prediction of the congestion level (DBSCAN).

ID	Period (DD/MM/YY)	KNN	R. Forest
33	09/01/2019 – 16/01/2019	LOW	NOISE
34	17/01/2019 – 24/01/2019	MEDIUM	MEDIUM
35	25/01/2019 – 31/01/2019	MEDIUM	MEDIUM
36	01/02/2019 – 08/02/2019	MEDIUM	MEDIUM
37	09/02/2019 – 16/02/2019	HIGH	HIGH
38	17/02/2019 – 24/02/2019	HIGH	MEDIUM
39	25/02/2019 – 31/02/2019	MEDIUM	MEDIUM
40	01/03/2019 – 08/03/2019	MEDIUM	MEDIUM
41	09/03/2019 – 16/03/2019	MEDIUM	MEDIUM
42	17/03/2019 – 24/03/2019	LOW	NOISE

The confusion matrix, precision, recall, and f1-score are the most commonly used metrics [42]. Confusion matrix C is such that C_i is equal to the number of observations known to be in group i and predicted to be in group j [43]. The general idea is to count the number of times instances of class x are classified as class y. Accuracy is the ratio of the total number of correct predictions and the total number of predictions [44]. Next, Table 5 and 6 present the confusion matrix and accuracy of the supervised learning algorithms, being "R" the actual value and "P" the predicted value.

Table 5: Confusion Matrix based on K-means.

Algorithm	R / P	High	Medium	Low	Accuracy
KNN	High	4	0	0	100%
	Medium	0	4	0	
	Low	0	0	2	
R. Forest	High	4	0	0	100%
	Medium	0	4	0	
	Low	0	0	2	

Table 6: Confusion Matrix based DBSCAN.

Algorithm	R / P	High	Medium	Low	Noise	Accuracy
KNN	High	1	0	0	0	70%
	Medium	1	5	0	0	
	Low	0	0	1	0	
	Noise	0	1	1	0	
R. Forest	High	1	0	0	0	80%
	Medium	0	6	0	0	
	Low	0	0	0	1	
	Noise	0	1	0	1	

Due to limited data availability, the accuracy of the supervised machine learning algorithms employed was biased. However, the continuous evaluation of the algorithms is crucial as the database grows. Since future instances have unknown target values, the metrics utilized will determine the accuracy of predicting new values. It's essential to understand the context and to select the appropriate methods carefully. The correct prediction of the congested periods could assist the port authorities in better planning and management.

5. Conclusions

Port congestion is an important issue from an economic and efficiency point of view. Because it results in longer waiting times and low service levels for vessels, it also contributes to the decrease in competitiveness and demand. Understanding the aspects that influence congestion is essential for port management. However, most studies performed with AIS data have focused on specific areas such as monitoring, tracking, and security of ships, accident prevention, including collision risks, noise levels, or ship emissions. Few kinds of research employing AIS data have explored marine traffic congestion.

To analyze the port congestion level of Rio de Janeiro were proposed three Port Congestion Indicators (PCIs). From January 2018 to April 2018 and September 2018 to March 2019, AIS data were collected and analyzed to calculate the geospatial algorithms: convex hull area, geolocation area, and average vessel proximity. The Port Congestion Indicators (PCIs), spatial concentration, spatial density, and average service time were formulated based on the geospatial algorithms previously calculated. Then, unsupervised (K-means and DBSCAN) and supervised (K-Nearest Neighbors and Random Forest) learning algorithms were employed to identify the congested periods and predict the congestion level of Rio de Janeiro port.

The K-means and DBSCAN algorithms were employed to cluster the normalized Port Congestion Indicators. The K-means algorithm was performed, setting $K = 3$ to aggregates the periods into three clusters. The DBSCAN algorithm was performed setting the $MinPts$ ($K = 2$). Once the $MinPts$ value is set, were calculated the Best Eps Value (the maximum point of curvature or most significant slope on a graph) for each scenario. Table 2 identified the periods which were considered as high congestion, for both algorithms, that could require some actions described in section 4.1.

The K-Nearest Neighbors and Random Forest were employed to predict the congestion levels of Rio de Janeiro port based on the previous classification of the K-means and DBSCAN algorithms. Given the approach used in machine learning, were employed a training subset about $70\% - 80\%$ and a test subset about $20\% - 30\%$. The x training subset contains the normalized attributes spatial concentration, spatial density, and average service time. The y training subset contains the labeled data classified by K-means and DBSCAN algorithms.

Due to limited data availability, the accuracy of the supervised machine learning algorithms employed was biased. However, the continuous evaluation of the algorithms is crucial as the database grows. The correct prediction of the congested periods could assist the port authorities.

However, it is essential to improve the proposed indicators for a reviewed advanced model. The proposed indicators are reactive and based on historical AIS data. Nonetheless, it can be applied provocatively to classify port congestion based on a real-time AIS data stream. Apache Spark may be required to fast and distributed engine for large-scale data processing. It provides distributed machine learning capabilities and can be reconfigured to enable real-time data processing. It might be interesting to compare different ports to start investigating why port congestion is occurring. Is the port overutilized, and should its operation be improved, or are there other reasons such as weather conditions or internal processes that need to be better accounted for?

Port congestion is increasing costs for shippers and importers at Rio de Janeiro port. Few studies employing AIS data to analyze port congestion were developed. Thus it is a contribution that can be useful, and the methodology can be applied to other ports. These surveys expand knowledge in the field since the quantification of port congestion provides port authorities resources to manage better and plan, improve port logistics operations, and reduce costs.

References

[1] ABUALHAOL, I. et al. *Mining port congestion indicators from big AIS data.* In: 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018. p. 1-8.

[2] Rosssi, M. (16 de 12 de 2019). *Navios esperam 10 dias para atracar em Santos e prejuízo ultrapassa US$ 10 milhões. G1 Santos.* https://g1.globo.com. Accessed: 20/01/2020.

[3] SAEED, N. et al. *Governance mode for port congestion mitigation: A transaction cost perspective.* NETNOMICS: Economic Research and Electronic Networking, v. 19, n. 3, pp. 159-178, 2018.

[4] ZHANG, L. et al. *Big AIS data based spatial-temporal analyses of ship traffic in Singapore port waters.* Transportation Research Part E: Logistics and Transportation Review, v. 129, pp. 287–304, 2019.

[5] MENG, Q. et al. *analysis with automatic identification system data of vessel traffic characteristics in the Singapore strait.* Transportation Research Record, v. 2426, n. 1, pp. 33-43, 2014.

[6] SHELMERDINE, R. L. Teasing out the detail: How our understanding of marine AIS data can better inform industries, developments, and planning", Marine Policy, v. 54, pp. 17{25, 2015.

[7] CRAIGHEAD, C. et al. *The severity of supply chain disruptions: design characteristics and mitigation capabilities*. Decision Sciences, v. 38, n. 1, pp. 131-156, 2007.

[8] MARTINS, V. Dynamic *Port Congestion Indicators - Case Study of the Ports of Rio de Janeiro and Santos.* Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Oceânica, COPPE, da Universidade Federal do Rio de Janeiro. Maio de 2021.

[9] MARTINS, Vinícius Barreto et al. *A Dynamic Port Congestion Indicator – A Case Study of the Port of Rio de Janeiro*. 28º International Congress on Waterborne Transportation, Shipbuilding and Offshore Constructions. Rio de Janeiro, October, 2020.

[10] POWELL, G. *Beginning database design.* John Wiley & Sons, 2006.

[11] CDRJ. *Porto do Rio de Janeiro - Características Gerais.* https://www.portosrio.gov.br. Accessed: 28/09/2020.

[12] TALEND. *What is Data Processing?* https://www.talend.com/Resources. Accessed: 14/09/2020.

[13] PEDA. *Data processing*. https://peda.net. Accessed: 14/09/2020.

[14] TEAM, P. *The 6 Stages of Data Processing Cycle.* https://medium.com/peerxp/the-6-stages-of-data-processing-cycle. Accessed: 14/09/2020.

[15] CODD, E. F. *Relational database: A practical foundation for productivity.* In: Readings in Artificial Intelligence and Databases, Elsevier, pp. 60-68, 1989.

[16] HARRINGTON, J. L. *Relational database design and implementation*. Morgan Kaufmann, 2016.

[17] HALPIN, T., MORGAN, T. *Information modeling and relational databases*. Morgan Kaufmann, 2010.

[18] KNIGHT, M. *What is Structured Query Language (SQL)?* https://www.dataversity.net.
Accessed: 14/02/2021.

[19] BECKER, R. *Structured Query Language (SQL)*. https://www.techopedia.com/definition.
Accessed: 14/02/2021.

[20] CODECADEMY. *What is a Relational Database Management System?* www.codecademy.com.
Accessed: 14/02/2021.

[21] MAPPS. *What is geospatial?* https://www.mapps.org/page/WhatisGeospatial.
Accessed: 09/10/2020.

[22] AAAS. *What are geospatial technologies?* https://www.aaas.org/programs/scientific.
Accessed: 09/10/2020

[23] BARBER, C. et al. The Quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software (TOMS), v. 22, n. 4, pp. 469-483, 1996.

[24] MUCKE, E. *Computing Prescriptions: Quickhull: Computing Convex Hulls Quickly*. Computing in Science & Engineering, v. 11, n. 5, pp. 54-57, 2009.

[25] XIANG, W. *An Efficient Location Privacy Preserving Model based on Geo-hash*. In: 2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), pp. 1–5. IEEE, 2019.

[26] JANAKIEV, N. *Calculate Distance Between GPS Points in Python*. https://janakiev.com.
Accessed:03/08/2020.

[27] THEOBALD, O. *Machine learning for absolute beginners.* 2017.

[28] ALPAYDIN, E. *Introduction to machine learning.* MIT press, 2020.

[29] GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* O'Reilly Media, 2019.

[30] GHAHRAMANI, Z. *Unsupervised learning.* In: Summer School on Machine Learning. Springer, Berlin, Heidelberg, 2003. p. 72-112.

[31] PATEL, A. *Hands-on unsupervised learning using Python: how to build applied machine learning solutions from unlabeled data.* O'Reilly Media, 2019.

[32] BERRY, M. W., MOHAMED, A., YAP, B. W. Supervised and Unsupervised Learning for Data Science. Springer, 2019.

[33] INDIA, A. *Types of Classification Algorithms.* https://analyticsindiamag.com/typesclassification.
Accessed: 20/01/2021.

[34] MONKEYLEARN. *Classification Algorithms in Machine Learning: How They Work.* https://monkeylearn.com.
Accessed: 20/01/2021.

[35] PETERSON, L. *K-nearest neighbor.* Scholarpedia, v. 4, n. 2, p. 1883, 2009.

[36] BREIMAN, L. *Random forests.* Machine learning, v. 45, n. 1, p. 5-32, 2001.

[37] PARKER, B. *Reduce Port Congestion by Expanding Container 'Free-Flow' Systems.* https://www.supplychainbrain.com/articles.
Accessed: 04/03/2021

[38] XCHANGE. *Port Congestion an Industry threat.* https://container-xchange.com/blog/port-congestion.
Accessed: 04/03/2021.

[39] BLANCHARD, D. *Five Strategies to Avoid Port Congestion.* https://www.industryweek.com.
Accessed: 04/03/2021.

[40] SCIKIT. *K-Nearest Neighbor*. https://scikit-learn.org/stable/modules/neighbors.html.
Accessed: 25/01/2021.

[41] NAVLANI, A. *Understanding Random Forests Classifiers in Python*. https://www.datacamp.com/ community/tutorials. Accessed: 04/02/2021.

[42] ZOLTAN, C. *KNN in Python.* https://towardsdatascience.com/knn-in-python.
Accessed: 25/01/2021.

[43] SCIKIT-LEARN. *Scikit-Learn - Confusion Matrix*. https://scikit-learn.org/stable/modules/generated
Accessed: 05/03/2021.

[44] FOR GEEKS, G. *Confusion Matrix in Machine Learning*. https://www.geeksforgeeks.org/confusion-matrix
Accessed:05/03/2021.