Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 238 (2024) 216–223

Procedia Computer Science

www.elsevier.com/locate/procedia

The 15th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 23-25, 2024, Hasselt, Belgium

Elucidating US Import Supply Chain Dynamics

Nikolay Aristov[1][a], Ziyan Li[1][a], Thomas Koch[a], Elenna R. Dugundji[a,*]

[a]*Center for Transportation and Logistics, Massachusetts Institute of Technology, 1 Amherst St., Cambridge, MA, 02142, USA*

Abstract

To enhance understanding of congestion points at ports and provide visibility into the incoming goods flow into the USA, this study focuses on maritime ports, using the Port of Boston and New York/New Jersey as case studies. Based on the Automatic Information System (AIS) data, we aim to develop predictive models for port congestion status and the Estimated Time of Arrival (ETA) of container ships. Additionally, we analyze historical commodity flow data to forecast future values, weights, volumes and categories based on Harmonized System (HS) codes. Employing quantitative AIS data analysis provides insights into port congestion dynamics and commodity flow trends, indicating the potential to improve the accuracy of ETA, port management and logistics visibility. This study contributes to both theoretical and practical applications in maritime logistics.

Keywords: Automatic Information System (AIS); Port congestion status; Estimated Time of Arrival (ETA); Density-based spatial clustering of applications with noise (DBSCAN); Spatial Temporal Graph Neural Network (STGNN); Imported commodities; Boosted Hybrid

1. Introduction

To increase the resilience of the nationwide supply chain network, it is necessary to have a better understanding of the processes at the most upstream part of goods flow in the USA – the ocean ports. This understanding will give information about the estimated arrival time of goods, their category, quantity, and routing. The information obtained may then be used by stakeholders for better planning of transportation, allocation of resources, ordering, and sourcing of goods.

The initial problem could be split into two major parts. The first part of the problem should be the prediction of the Estimated Time of Arrival (ETA) of container ships to the port, which will give us an estimation of the goods import dates into the USA. In this part, we will deeply study the Automatic Information System (AIS) data, and select and compare different statistical or machine learning models to predict ETA. [15] The International Maritime Organization's International Convention for the Safety of Life at Sea requires AIS to be fitted aboard international

[1] These authors have equal contribution
[*] Corresponding author. Tel.: +1-617-258-6048
E-mail address: e.r.dugundji@mit.edu

10.1016/j.procs.2024.06.018

Field Name	Description
MMSI	Maritime Mobile Service Identity, unique nine-digit identification number for each vessel
BaseDateTime	Date and time of the AIS signal
LAT, LON	Geographical coordinates of the vessel
SOG	Speed over ground in knots
Draught	Draught of ship
COG	Course over ground in degrees from true north
IMO	IMO ship identification number, a unique and permanent seven-digit identification number
Static data	Length, width, draft, etc.

Table 1. information from Automatic Information System (AIS) data

voyaging ships with 300 or more gross tonnage, and all passenger ships, regardless of size. [6] As shown in table 1, the AIS collects at regular intervals vessel data throughout the maritime sailing in radio frequency, improving safety and traceability in global ocean logistics.

Besides the usage of the historical AIS data for ETA estimation, we have to consider other exogenous factors, such as delays at the transshipment ports, congestion near the port, and supply chain disruptions during pandemics or strikes, which may affect the accuracy of ETA prediction. We will research the influence of these factors in future studies.

The second part of the problem is to analyze the goods flowing through the port. There could be level, trend, and seasonality in the underlying model of goods flow; knowing these parameters and their interrelationships will provide good visibility for all the stakeholders in the downstream goods flow.

Our study will use the Port of Boston (operated by the Massachusetts Port Authority (Massport)) and New York/New Jersey as case studies, aiming to develop a spatial-temporal analysis of the supply chain dynamics of global ocean logistics networks. Additionally, we will build a predictive model of the commodities flow trends imported through the port of Boston.

2. Literature review

In recent research projects, several studies have delved into predicting vessels' behaviors, addressing challenges related to congestion and traffic flow, and optimizing port operations. Also, besides statistical models, more advanced methodologies have been applied in transportation research, such as Neural Networks and Transformer, showing promising results in predicting ETA. In this section, literature regarding these areas will be discussed.

2.1. ETA Prediction

As described in Section 1, the problem of ETA prediction could be separated into several parts. We start with papers regarding trajectory estimation and continue with works on the prediction of traffic congestion and modeling of the port's operations in the immediate vicinity of the port. We also research articles predicting the ETA of other modes of transportation, such as trains or truckloads, which also give us insights on how to make predictions using different methods.

2.1.1. Traffic Congestion

The ability to find congestion points and predict traffic flow is crucial for accurate ETA predictions. Analyzing historical AIS data allows us to identify anchor and berth areas of the port by utilizing a specially developed algorithm based on Density-based spatial clustering of applications with noise (DBSCAN) method [2]. This algorithm was tested on eight ports with complicated geographic features and could be an appropriate starting point for the analysis of port ecosystems. The algorithm itself could also be used for monitoring congestion data at the specified ports.

[16] showed that eXtreme Gradient Boosting (XGBoost) and Shapley Additive Explanation (SHAP) could be used to predict port congestion status and improve the prediction accuracy of time spent in port. It was also stated that for

predicting the traffic flow rate, the XGBoost algorithm had the lowest error for hour-ahead forecasts in comparison to Holt-Winter, Transformer, and Graph Neural Network (GNNs) [3].

[9] used Spatio-Temporal Adaptive Graph Convolutional Networks (STAGCN) to extract the properties of the road network topology graph. First, the authors captured the structure of the road network traffic by using an adaptive graph generation block, built an adaptive road network topology graph, and then fed the result to capture spatial-temporal features of the traffic data by utilizing spatial-temporal convolution blocks. This work tested the approach on publicly available datasets for freeway traffic and claimed that prediction accuracy outperformed modern baseline methods. However, the authors stated that STAGCN has limitations as it requires two features: traffic flow and traffic speed.

Another approach for predicting congestion combines several methods in sequence. First, Variational Mode Decomposition (VMD) analyzes traffic flow and returns the corresponding model components based on the number of modes (which is a parameter of the method). Next, the result is used by the Extreme Learning Machine (ELM) to predict the traffic flow, and then the Whale Optimization Algorithm (WOA) is applied to optimize the internal ship traffic flow parameters of the network. The result of this VMD-ELM-WOA then used as an input for a Fuzzy c-means (FCM) to classify the level of ship traffic congestion [4].

One more approach is to look at congestion from the perspective of port service capabilities. One of the studies proposed a ship traffic model based on ship size, arrival time and service time for use in terminal design [14].

We will utilize the algorithm DBSCAN as it showed good results in clustering and detecting congestion points. We will detect the status of the port by applying XGBoost and SHAP and compare this approach to VMD-ELM-WOA. We will also consider findings from the study by [14].

2.1.2. ETA Prediction

Many research papers regarding the prediction of ETA in maritime were published recently, and this topic is highly developed currently. Most of the studies concentrate on the port area. The simplest way to predict ETA is to derive it from trajectory and speed over ground [11, 1]. However, more sophisticated methods could be used for ETA prediction, such as CNNs and RNNs [5] or Hidden Markov Model (HMM) [13]

Predicted ETA could be used as input for machine learning models for port call optimization [7] and berth scheduling [8] as well as for good flow prediction models.

We will utilize and compare the discussed approaches for ETA prediction in our work.

2.2. Analysis of Commodities

Having a comprehensive understanding and prediction of incoming commodities is important for the operation and development of a seaport, so that the port can be prepared for berth operation based on the incoming commodities' categories and volumes. Several forecasting models have been used to gain a deeper insight into the imported commodities, for example, regression, time series statistical models, and even neural networks [12]. With limited data, time series models such as Autoregressive Integrated Moving Average (ARIMA) are common methods to provide a promising prediction of the portfolio of incoming commodities.

Also, another aspect that has been considered is that the value-weight ratio from historical data on imported commodities could be connected with trade routes to identify which ocean trade routes help the seaport make the best allocation of its equipment and human resources. [10]

We will use and compare the discussed approaches for commodity trends.

3. Methodologies

The prediction of the estimated time of arrival of the vessel could be split into several parts based on the routing. The first part is the estimation of trajectories up to arrival in the waiting zone at the Port of Boston. We will compare several statistical and machine learning models based on historical data and routes, as well as projections of congestion at reference points along the way, to find the model providing the best estimation. For this part, we will analyze the historical routes of container ships to identify average times and congestion points, i.e., places with the most time spent. Based on the congestion, we will come up with a model for the prediction of ETA using both identified factors and congestion points. We will incorporate prediction models for the time spent at the congestion points into the overall ETA prediction model.

In the next step, we will use the output of estimated times of arrival at port waiting zones as the inputs for a model of the port operations. By using historical data, we will try to find factors affecting the traffic flow rate in the port area and build a prediction model of traffic flow. Apart from the traffic, we will also identify patterns of container ship trajectories based on vessel direction, entrance point and other parameters that could affect the ETA of the vessels to the anchorage and berth areas. We will build several machine learning models and compare them to statistical models.

In the last step, we will describe the imported commodities' time series features and use the predicted ETA as a parameter for predictive models of commodities' flow through the Port of Boston.

3.1. *Prediction Models of the ETA of Container Ships*

In this section, we will start with the strategies we used for data handling.

3.1.1. *Data Handling*

We filtered AIS data by container ships in the vicinity of the coastline of the USA for 2015-2023 years. We also calculated number of all vessels in the harbor area and container ships near the berth and in the queue on daily basis. At this moment we don't consider any exogenous factors for our model as well as port utilization and operation time from any other source except historical data received from AIS data.

3.1.2. *DBSCAN*

We performed initial clustering for the whole data set to figure out possible points of interest. We were unable to identify any additional points of interest except ports and their anchorage areas. However, our analysis is missing data for areas like the Panama Canal that could affect the ETA of ships. For narrowing our search, we filtered data by speed over ground, assuming that ships in anchorage and berth areas are spending some time without moving. Applying the DBSCAN algorithm allowed us to identify berths and anchorage areas for Massport and other ports in the USA. The anchorage areas will be used in the next steps to build the prediction model of congestion times. We are satisfied with the outcomes produced by the algorithm for the initial point of interest detection. The results provided by the algorithm are deemed satisfactory, considering its effectiveness in addressing the primary objectives of our study. As this phase serves as the initial step in our research, we have opted not to investigate additional algorithms in this area at the current stage. The obtained results sufficiently fulfill the requirements for the preliminary analysis, allowing us to focus our efforts on the subsequent stages of our research.

3.2. *Commodities*

3.2.1. *Data Handling*

We began the analysis for commodities imported into the Port of Boston from the importation data-set by 2-digit Harmonized System (HS) code from 2003 to 2023. Filtered by containerized value and weight, we selected the top 5 commodities, respectively, in order to catch the trend, seasonality and other nonlinear relationships and build predicting models.

3.2.2. *Statistical Models*

Based on the trend and seasonality shown in the figures below, we decided to first apply the statistical models Holt-Winter and ARIMA. They are both time series models to catch the trend and seasonality, while ARIMA performs better with data showing a non-stationary trend.

3.2.3. *Boosted Hybrid*

Boosted Hybrid is a hybrid model that combines a linear regression model that catches normal variables infected with the target values and trains the residuals from the first model with other exogenous or uncommon variables to catch the peaks or valleys in the data.

According to the highly oscillated historical data of imported commodities, we see a great possibility to apply boosted hybrids to predict the imported commodities' value and weight more accurately.

For the linear regression variables, we considered lags and holiday issues, took Christmas as a one-hot feature for removing seasonality from the actual data. We plotted the unseasoned data by its autocorrelation, partial autocorrelation and lags.

After calculating the residuals from the first model, we applied the random forest model to train the residuals by features such as month of year and holiday months.

4. Initial Analysis and Results

4.1. Congestion Points

We performed initial clustering for the whole AIS data set to figure out possible points of interest. At first, the clustering was done with a radius of 10 km, and we were able to identify areas of interest. By applying our knowledge of port specifics, we were also able to identify problems with AIS data, for example, wrong MMSI identification. We also saw that we should reduce the radius for detection of port terminals in complicated areas. By decreasing this number to 2 km, we were able to identify all terminals in New York / New Jersey port, but we still need more precise clustering for ports with complicated berth geometry and narrow separation of terminals (for example, Los Angeles / Long Beach). However, analysis identified for us all ports with container terminals in the USA and congestion points like, for example, the bridge near Annapolis on the way to the port of Baltimore. We could also clearly see increasing and decreasing congestion in ports over time.

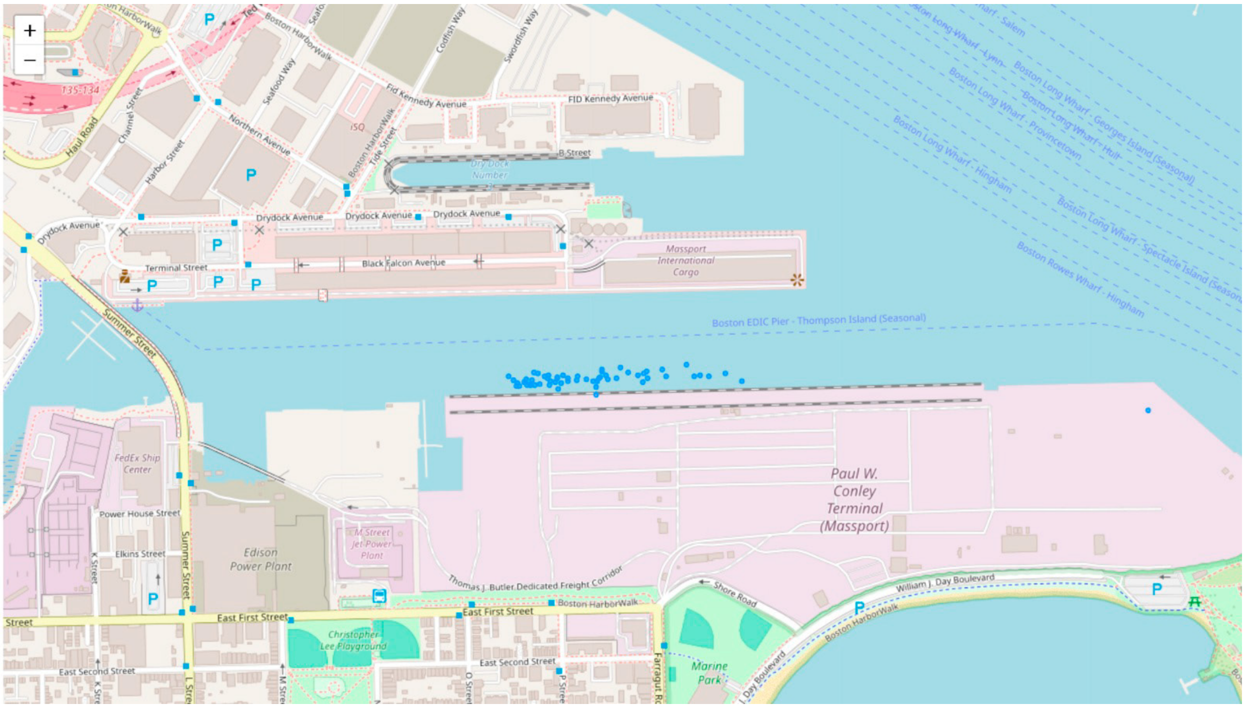
+
−
Northern Avenue
Drydock Avenue
Black Falcon Avenue
Terminal Street
FID Kennedy Avenue
B Street
Massport International Cargo
Boston EDIC Pier - Thompson Island (Seasonal)
Boston Long Wharf - Hingham
Boston Long Wharf - Provincetown
Boston Long Wharf - Georges Island (Seasonal)
Boston Long Wharf - Spectacle Island (Seasonal)
Boston Rowes Wharf - Hingham
Summer Street
FedEx Ship Center
Power House Street
Elkins Street
Edison Power Plant
M Street Jet Power Plant
Thomas J. Butler Dedicated Freight Corridor
Paul W. Conley Terminal (Massport)
East First Street
East Second Street
Christopher Lee Playground
Shore Road
William J. Day Boulevard
Boston HarborWalk
Marine Park
Farragut Road

Fig. 1. Port of Boston Berth area

By performing statistical analysis of AIS data for the Boston Port area (berths and anchorage areas) we observed no seasonal or time effect. We confirmed with port authorities that the only reason, except some non-related to port (like technical or rescheduling), for ship timing in the wait area is traffic inside the harbor.

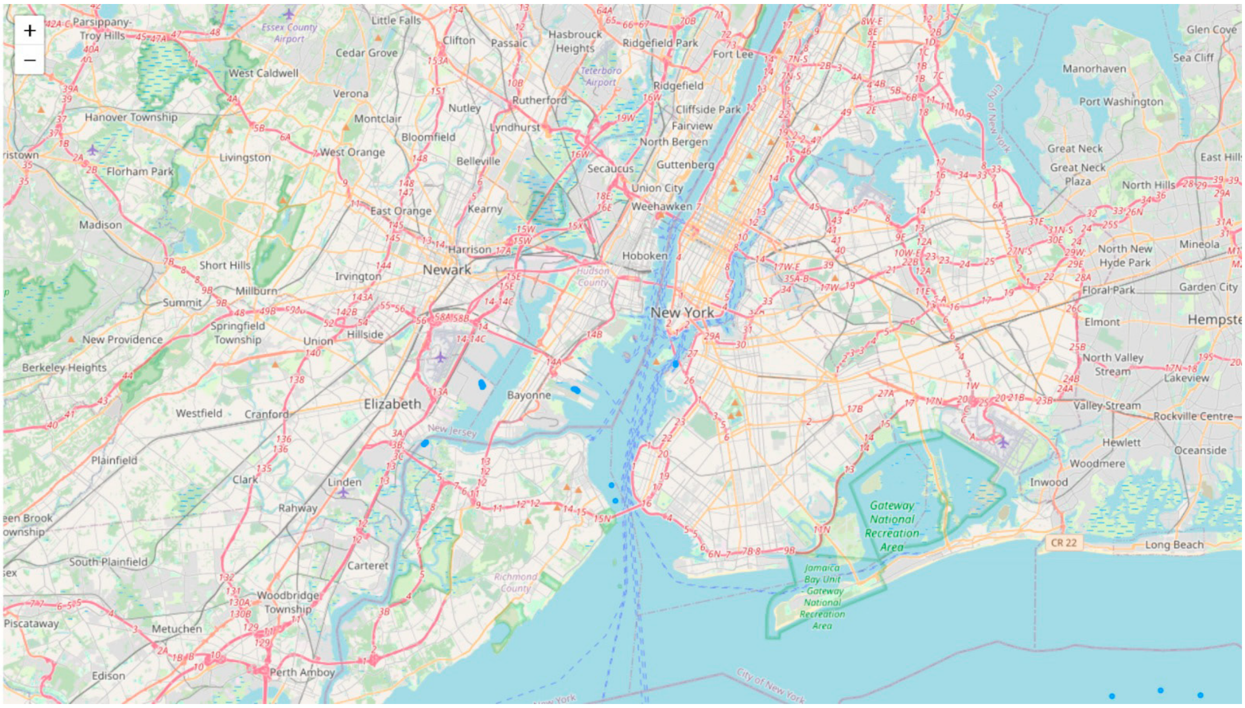

Fig. 2. New York / New Jersey terminal identification

	Holt Winter	ARIMA	boosted hybrid
MAPE for predicting values	59.18%	69.12%	40.17%

Table 2. Comparison of MAPE for testing forecasts among different models

4.2. Commodities

Based on the import data by 2-digit HS code from 2003 to 2023, we first took a look at the five most valued commodities imported by Massport, and the most valuable commodity during the past 20 years was HS code 84 (Nuclear Reactors, Boilers, Machinery, Etc.). Performed analysis showed that commodity 84 (Nuclear Reactors, Boilers, Machinery, Etc.) has a high correlation with commodity 39 (Plastics And Articles Thereof) with a pairwise correlation 0.88. So, we decided to aggregate these two commodities to see a general feature.

4.2.1. Statistical Models

We employed Holt-Winter and Autoregressive Integrated Moving Average (ARIMA) methodologies to analyze the time series data. The data was partitioned into 80% for training and 20% for testing.

4.2.2. Boosted Hybrid

As described in Section 3, a two-tiered hybrid modeling approach comprises linear regression for time steps and lagged variables, while a random forest regressor was employed to train on the residuals derived from the first model. As depicted in Figure 3, the boosted hybrid model yield relatively more accurate predictions for the values of commodities 84 and 39.

On the other hand, we also analyzed the most weighted commodities: 22 Beverages, Spirits And Vinegar. With the same approach, liquids' weight showed a more stable trend and seasonality. Figure 4 shows the boosted hybrid model for the weight of beverages.

As a result, the boosted hybrid performed better than other statistical models based on error term of Mean Absolute Percentage Error (MAPE) for testing forecasts.

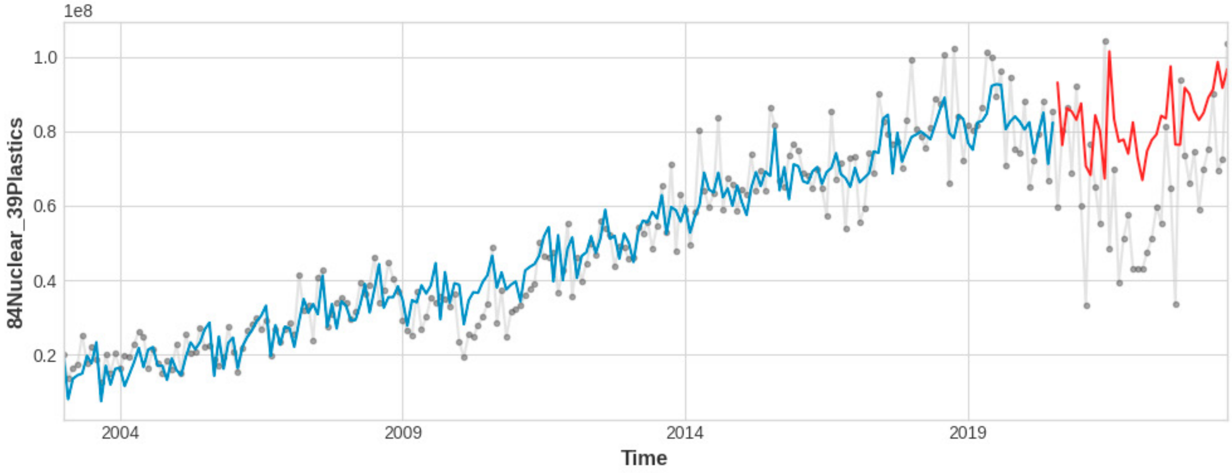

Fig. 3. Predictive Model for aggregated value of commodity 84 and 39

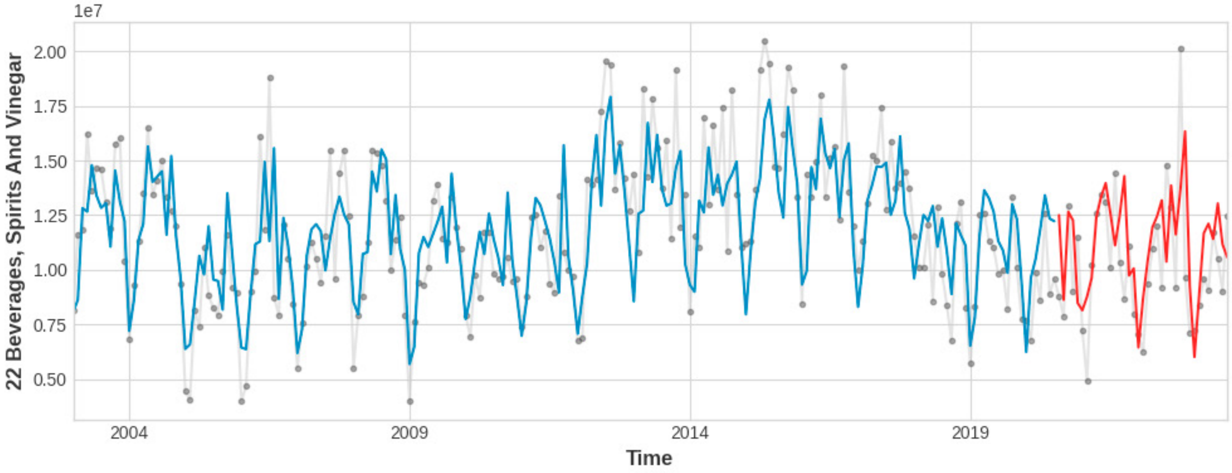

Fig. 4. Predictive Model for weight of commodity 22

5. Conclusion

5.1. *DBSCAN*

DBSCAN algorithm showed very robust results in clustering AIS data and detecting congestion points. However, additional analysis and application of industry knowledge are required for the obtained data. The results we received allowed us to narrow areas of interest and monitor traffic rates in them.

Building the congestion model for New York / New Jersey port will allow better prediction of the ETA of all connected (consequent in ship's routes) ports. It is required to analyze the quality of operations inside the port and harbor to build the model of a particular port.

5.2. *Commodities*

The boosted hybrid model provided a relatively accurate predictive data for both the value and weight of incoming commodities. Having an accurate prediction of incoming commodities can help not only port authorities but also all the stakeholders in the supply chain to be better prepared for berths and truckload capacity, or even better allocation in hinterland logistics.

6. Future Research

In the next steps, we are planning to build the congestion model for the New York / New Jersey port. We also plan to apply and test the model at the Los Angeles / Long Beach port. Then, we will build the GNNs model to predict all congestion statuses of ports and ETA. We also plan to compare this model to several other models, for example, statistical.

We will try to identify parameters affecting commodity flow over time by source region, destination port and economic indexes. This will allow us to identify inputs for additional scenarios for predicting the effects of possible interruptions or changes in congestion rates.

On the commodity side, based on current data, we could still inspect a big lap between actual data and predictive data of value around years 2021 and 2022. The dramatic drop could be a result of corruption in production and importation due to the pandemic. This indicates that we still need more exogenous variables to analyze the influence of the global economic environment or black swan incidents such as pandemics on incoming commodities.

Similar to the congestion analysis, it is also important to predict the commodity flow outside the Boston port area. The commodity flow cannot be treated without analyzing the whole incoming flow in the USA and the sequence of ports on route.

References

[1] Alessandrini, A., Mazzarella, F., and Vespe, M. (2018). Estimated time of arrival using historical vessel tracking data. *IEEE Transactions on Intelligent Transportation Systems*, 20(1):7–15.

[2] Bai, X., Ma, Z., Hou, Y., Li, Y., and Yang, D. (2023). A data-driven iterative multi-attribute clustering algorithm and its application in port congestion estimation. *IEEE Transactions on Intelligent Transportation Systems*.

[3] Belt, E. A. (2023). Analysis and short-term forecasting of traffic intensity: Exploring the impact of road maintenance. *Master's thesis*.

[4] Chen, Y., Huang, M., Song, K., Wang, T., et al. (2023). Prediction of ship traffic flow and congestion based on extreme learning machine with whale optimization algorithm and fuzzy c-means clustering. *Journal of Advanced Transportation*, 2023.

[5] El Mekkaoui, S., Benabbou, L., and Berrado, A. (2023). Deep learning models for vessel's ETA prediction: bulk ports perspective. *Flexible Services and Manufacturing Journal*, 35(1):5–28.

[6] International Marine Organization (2004). Ais transponders. https://www.imo.org/en/OurWork/Safety/Pages/AIS.aspx. Effective: 2004-12-31.

[7] Jahn, C. and Scheidweiler, T. (2018). Port call optimization by estimating ships' time of arrival. dynamics in logistics. In *Proceedings of the 6th International Conference LDIC*.

[8] Kolley, L., Rückert, N., Kastner, M., Jahn, C., and Fischer, K. (2023). Robust berth scheduling using machine learning for vessel arrival time prediction. *Flexible Services and Manufacturing Journal*, 35(1):29–69.

[9] Ma, Q., Sun, W., Gao, J., Ma, P., and Shi, M. (2023). Spatio-temporal adaptive graph convolutional networks for traffic flow forecasting. *IET Intelligent Transport Systems*, 17(4):691–703.

[10] Ong, G. P. and Sou, W. S. (2015). Modeling commodity value–weight trends between the United States and its trading partners. *Transportation Research Record*, 2477(1):93–105.

[11] Park, K., Sim, S., and Bae, H. (2021). Vessel estimated time of arrival prediction system based on a path-finding algorithm. *Maritime Transport Research*, 2:100012.

[12] Patil, G. R. and Sahu, P. K. (2016). Estimation of freight demand at Mumbai port using regression and time series models. *KSCE Journal of Civil Engineering*, 20:2022–2032.

[13] Waterbolk, M., Tump, J., Klaver, R., van der Woude, R., Velleman, D., Zuidema, J., Koch, T., and Dugundji, E. (2019). Detection of ships at mooring dolphins with hidden markov models. *Transportation Research Record*, 2673(4):439–447.

[14] Wawrzyniak, J., Drozdowski, M., and Sanlaville, É. (2022). A container ship traffic model for simulation studies. *International Journal of Applied Mathematics and Computer Science*, 32(4):537–552.

[15] Yang, D., Wu, L., Wang, S., Jia, H., and Li, K. X. (2019). How big data enriches maritime research–a critical review of Automatic Identification System (AIS) data applications. *Transport Reviews*, 39(6):755–773.

[16] Zhang, T., Yin, J., Wang, X., and Min, J. (2023). Prediction of container port congestion status and its impact on ship's time in port based on AIS data. *Maritime Policy & Management*, pages 1–29.