Industrial & Applications Paper

Patterns of Life : Global Inventory for maritime mobility patterns

Giannis Spiliopoulos
MarineTraffic (Kpler), Research Labs
Athens, Greece
Intelligent Transportation Systems Lab, University of the Aegean
Ermoupolis, Syros, Greece
gspiliopoulos@aegean.gr

Marios Vodas
MarineTraffic (Kpler), Research Labs
Athens, Greece
mvodas@kpler.com

Georgios Grigoropoulos
MarineTraffic (Kpler), Research Labs
Athens, Greece
ggrigoropoulos@kpler.com

Konstantina Bereta
MarineTraffic (Kpler), Research Labs
Athens, Greece
kbereta@kpler.com

Dimitris Zissis
Intelligent Transportation Systems Lab, University of the Aegean
Ermoupolis, Syros, Greece
dzissis@aegean.gr

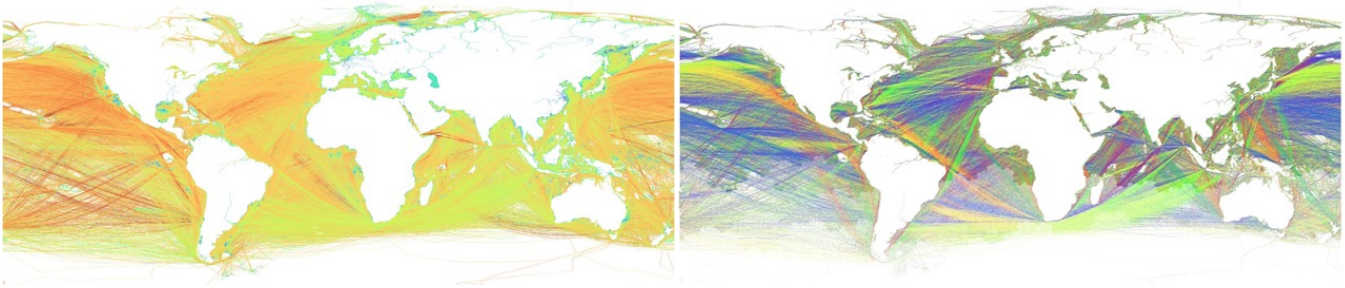

Figure 1: Patterns of Life for global vessel traffic of 2022 depicted in 7.3 million hexagonal cells of h3 resolution 6. Left: The average speed of 2022 for commercial fleet, blue is low speed and red is high. Right: Respectively the average course, green is North direction, red is south, blue is east and yellow is west.

ABSTRACT

More than 70% [22] of the global trade transportation is conducted by sea, through maritime sea lanes. Unlike the well defined global land transportation network that consists of roads and railways, the maritime equivalent consists of port connections and is vaguely defined by marine charts' guidelines, constraints and common sea routes. By definition, the port of origin and port of destination are well defined locations. However, the routes that the vessels follow in between are not strictly defined. Local conditions, such as the weather or traffic congestion, vessel-specific characteristics or other external conditions also affect the route choice and planning. The understanding of the typical behaviour of vessels sailing across the globe is crucial for the monitoring of the global logistic chain. A challenging data mining task is that of transforming the huge amounts of vessel tracking data currently available by the maritime industry data providers (such as MarineTraffic (Kpler)), into a descriptive and compact data model, that can be used for identifying the underlying relationships or patterns. In this work, we present a data-driven grid-based methodology that leverages big data distributed techniques, for extracting vessel mobility patterns on a global scale. The existence of a global inventory for the typical vessel behaviour patterns is crucial for the improved visibility of the global logistic chain and the timely identification of abnormal behaviour.

1 INTRODUCTION

In the last 4 years, maritime transportation has faced significant disruptions at a global scale [26]. Therefore, there is an increasing interest on improving its visibility and understanding by different stakeholders [5, 23]. Early in 2020, the COVID-19 pandemic measures that came into effect in some of the most important ports, significantly decreased their operational capabilities, effectively disrupting the global logistic chain, while segments of the passenger fleet faced the complete shutdown of their operations [14]. Later in 2021, a large container vessel blocked the Suez Canal for more than two weeks forcing vessel traffic towards the Mediterranean Sea to either re-route around Cape of Hope adding more than 7000 miles to their journeys [24] or to wait until the blockage was resolved. In the post COVID-19 era, innovative data solutions and advanced digital tools are required to meet the increased needs for efficiency, visibility and predictability of maritime operations [25] and to contain the repercussions of disruptive events.

The contribution of this work is described as follows: We present a scalable knowledge extraction methodology that combines trajectory data mining techniques and information fusion to generate

an inventory of regional mobility patterns that extends worldwide by extracting the implicit information from the Automatic Identification System (AIS) data. Through centralized vessel tracking systems that exploit AIS data it is possible to observe the evolution of incidents in real time. Key maritime stakeholders such as local authorities, terminal operation and maritime experts have nowadays the tools to timely detect incidents that occur within their responsibility areas. Nevertheless, they frequently find themselves inundated by the sheer magnitude of the data and their complexity, more than 400K vessels are tracked worldwide producing more than a billion of AIS records each day according to [9]. The approach described in this paper addresses the challenge to transform the substantial volumes of AIS tracking data into a descriptive and compact data model, that can be used in a computationally efficient way to identify underlying relationships or patterns.

Towards this goal, we present a data-driven grid based methodology that leverages big data distributed processing of a global AIS dataset to extract the local patterns and behaviours of vessels with respect the origin and destination port pair and the market segment each vessel belongs to. Our methodology consists of several steps that transform AIS messages to cell summaries of mobility features. We utilize the H3[1] spatial index, that belongs to the family of Geodesic Global discrete System (GGDS) [19] to organize AIS messages, which are part of different voyages of a specific port-to-port connection into local groups. As soon as the groups are identified, we calculate aggregates that can be used to define the distribution of mobility feature values in each cell. Additionally, we measure the popularity of transitions to nearby cells collectively since the order of messages per vessel is preserved. We dynamically collect all statistics for different combinations of port connections per vessel type based on vessel presence in each cell, and thus we organize this information into an easily accessible global inventory. Stakeholders can retrieve the historical statistical summary for each cell area, as well as the most frequent direct cell transition per market and port connections by querying for a specific location. This work is organised as follows. In Section 2, we present the related literature. In Section 3, we describe in detail the dataset, the spatial index and the execution framework, as well as the processing steps of our methodology. In Section 4, we present the compression achieved and demonstrate the added value of our approach on three important maritime use cases. Finally, in Section 5, we present the conclusions and the next steps for this work.

2 RELATED WORK

The key role of maritime transportation for the global trade, the availability of global scale sensor data, and the need for novel intelligent systems and processes that denote the first steps towards the fourth industrial revolution [17] has led to the increased interest of the scientific community to discover the latent information, hidden in mobility data sources such as the AIS. The analysis of AIS mobility data, falls under the "Trajectory Data Mining" (TDM) research topic. The term TDM is traditionally used, to describe the knowledge extraction from labelled and unlabelled data points that originate from moving objects. Trajectory data mining processes typically consists of several mining tasks and techniques, for an extended literature review of those we refer the reader to the survey [31]. As a plethora of different approaches exist that cannot be covered in this section, we focus on topics closely related to this study and not to the trajectory data mining in general. Thus, we present related work targeted to vessel route

The main data mining challenge in this case is, creating a descriptive model which identifies patterns of common trajectories. Clustering, summarization, association rules, and sequence discovery are usually viewed as appropriate methods in the related literature.

A common approach relies on unsupervised clustering techniques that organise data points or trajectory segments in groups. The *DBSCAN* [4] and *OPTICS* [1] algorithms are two well known density based clustering approaches that have been used extensively in the scientific literature. Both methods are used to identify dense concentrations of points that match certain criteria, filtering out the corresponding low density areas. In [29], authors organize AIS trajectories in a trip semantic objects (STSO) and the use density based clustering on points OPTICS to identify clusters of way-points and stops. Identified clusters are then organised in graph structure that represents the route network. Authors in [15], introduced the Route Extraction and Anomaly Detection (TREAD) methodology, that identifies without prior knowledge data points of importance and utilises an incremental-DBSCAN approach to organise them in representative routes. In [27] the *DBSCAN* algorithm is selected to cluster turning points of vessel and an artificial neural network is used to learn their connections. Authors in [18] exploit *OPTICS* capabilities for the definition of a route network and combine it with a probabilistic approach to define its boundaries. Additionally another density based method is presented in [10], where authors combine the kernel density estimation (KDE) with image processing techniques to extract the exterior boundaries high density areas detected by *KDE*. In an different direction from density based algorithms, but yet within the unsupervised clustering family of algorithms, the K-means algorithm in conjunction with map and reduce techniques are selected in [32]. In this work data points are partitioned based on specific journey and vessel semantics (i.e, map phase) and then apply the k-means on each partition, routes are modeled as the set of convex hulls of the identified clusters. In the approach introduced in [28], the authors introduce an unsupervised data mining technique that uses a density-based strategy to analyze vessels' trajectories and identify patterns in AIS data. In this work, DBSCAN is used to cluster AIS data.

Trajectory based clustering is another evolving approach for vessel proposed in several seminal works. In this approach, either segments or entire trajectories are organised in groups based on trajectory similarity criteria. In [11], authors use a multi-level trajectory clustering that is build upon k-means and dynamic time wrapping (*DTW*) similarity measure. In [30], the authors explore the effectiveness of the symmetrized segment-path short (SSPD) metric in trajectory clustering for vessel trajectories. A new trajectory clustering algorithm, named SPTCLUST-II was proposed in [3] for clustering vessel trajectories. The authors in this paper proposed a methodology that includes trajectory segmentation, clustering and route extraction. The approach described in [8] includes an unsupervised technique for trajectory extraction, compression and clustering. The compression task is based on the

[1] https://www.uber.com/blog/h3/

Manifold-blurring Mean-Shift algorithm and the Principal Component Analysis. The trajectory clustering task is based on the Longest Common Sub-sequence trajectory similarity measure and the agglomerative hierarchical clustering algorithm. In [12], the authors present a trajectory clustering technique that is based on a variation of DBSCAN and employs merge distance and multidimensional scaling to measure and represent trajectory similarity respectively.

Most of the techniques proposed in literature address the data mining procedure as a processing step for route extraction, estimation time of arrival calculation or anomaly detection. In most cases, existing literature focus on specific use cases, e.g, a local area, a specific route or a small selection of features. In preliminary work [13], we discuss the challenges of AIS dataset, and we compare performance of PostgreSQL[2] and MongoDB[3] in real work use cases. In [20], we highlight in a map-reduce knowledge extraction methodology, the sensitivity of DBSCAN against other clustering algorithms when applied on highly skewed, in terms of density, datasets such as global AIS ones.

The motivation behind the paper is that of transforming the huge amounts of geospatial data, into a descriptive and compact data model, that can be used for identifying the underlying relationships or patterns. We focus in the implementation of an efficient, data-driven methodology addressing scalability and generalization issues that have not been properly addressed before meeting real-world industrial requirements. In this case we build a model of normalcy that can then be used to identify any outliers from this e.g. Covid-19 or Suez Canal.

3 METHODOLOGY

In this section we present our method to construct a global maritime statistics inventory. Our method leverages big data technologies to process large amounts of vessel's trajectories data in parallel by extracting a large set of feature statistics from observations that are projected on the same cell of a predefined hexagonal grid (i.e., *H3*). Our methodology consists of several processing steps that are executed in sequence. Each step, internally capitalizes on the parallelization capabilities of Apache Spark, which is an open source unified engine for large scale analytics[4]. In summary, our methodology consists of the data cleaning process, the geofencing technique for reconstruction of port calls, the data enrichment step to bind positional reports with additional features, the feature extraction step that generates grid based summaries of the data. A visual representation of the methodology steps is shown in Figure 2. A detailed description of the dataset, the tools as well as more specific information for each step of our methodology follow in the next subsections.

3.1 Dataset description

3.1.1 AIS protocol. Since 2002 the International Maritime Organisation (IMO) has made compulsory for all vessels with a tonnage above 299 Gross Tonnage (GRT) to be equipped with an AIS class-A transceiver. The AIS was originally designed as a collision avoidance system. At its core, each AIS transceiver sends and receives positional reports (i.e., types 1-3 and 18) every few seconds via VHF signals that include information about each vessel's identity, location, course and speed. Since 2006 the lower-cost class B transceiver was introduced from the IMO allowing smaller vessels to use the AIS protocol with lower priority than the commercial fleet that strictly operates class-A transceivers [7]. The transmission rate of AIS ranges from 2 seconds, for fast moving vessels or maneuvering vessels equipped with a class-A transponder, and up to 3 minutes for anchored or moored vessels.

AIS messages can be captured from any vessel or onshore installation in VHF range, as long as the vessel is equipped with an AIS receiver. Regardless the case, the messages are decoded and the information received is presented in vessel navigational tools onboard or forwarded for further processing to databases and other applications. Since early 2010 there have been several global vessel tracking systems (*e.g* marinetraffic.com) that exploit available terrestrial (T-AIS) and satellite (S-AIS) AIS receivers network to gather, process and present global feeds of AIS messages in a centralized manner.Open AIS datasets covering smaller areas are available in [16, 21].

In this work, we use a global historical AIS dataset that consists of all positional reports that were received and archived by MarineTraffic (Kpler) for all vessels, throughout the year 2022. In addition, a vessel's static reports inventory was used to match each vessel present in the positional report dataset with its corresponding AIS vessel-type. The original dataset before any manipulation is 600GB large. We focus our analysis on vessels related to the logistic chain, so we filter out non commercial vessels' positional reports. After this preprocessing step the size of the dataset is decreased to 60GB, and it consists of positional reports of approximately 60 thousands vessels commercial vessels with a tonnage greater than 5000 GRT and equipped with class-A tranceiver.

Table 1: Data Used for Methodology

Description	Rows	Size
Commercial fleet positional reports	2.7 Billion	60GB
Vessel Static information	60 Thousand	few MB
Port Information	20 Thousand	few MB

3.2 Spatial index selection and execution framework

3.2.1 Spatial index selection. The spatial index and the respective grid system is at the core of our methodology and as soon as the following requirements are met, then it could be easily implemented with a different grid systems. The first requirement ensures the completeness of the solution as all vessels, regardless their location on earth, can be included in the calculations. Thus, the spatial index must be global (i.e., any location on earth can be represented within the grid) and that each cell must cover approximately the same area at a given resolution. The second requirement leads to interpretable

[2]https://www.postgresql.org/
[3]https://www.mongodb.com/
[4]https://spark.apache.org/

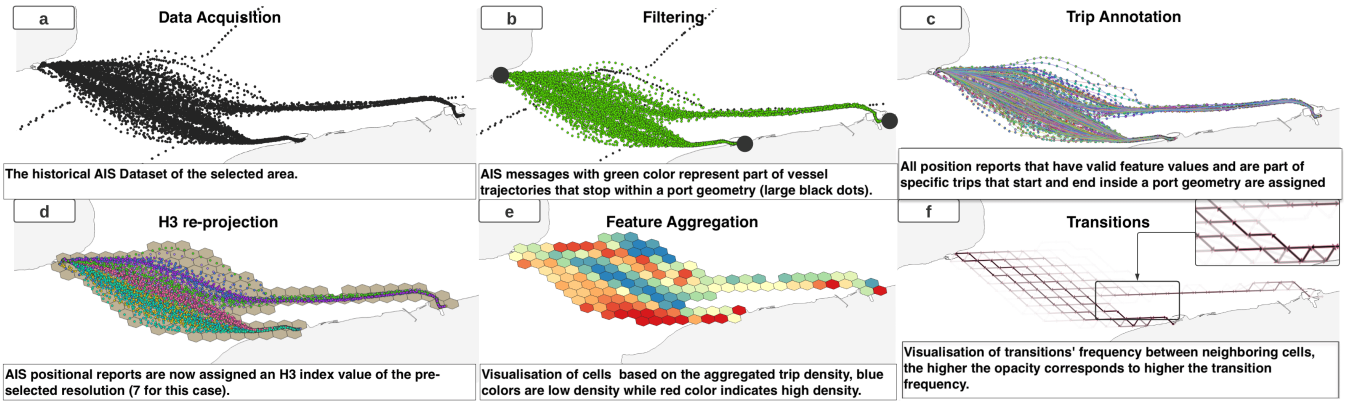

Figure 2: Pictorial representation of our methodology for a small AIS data-set in the area of English Channel.

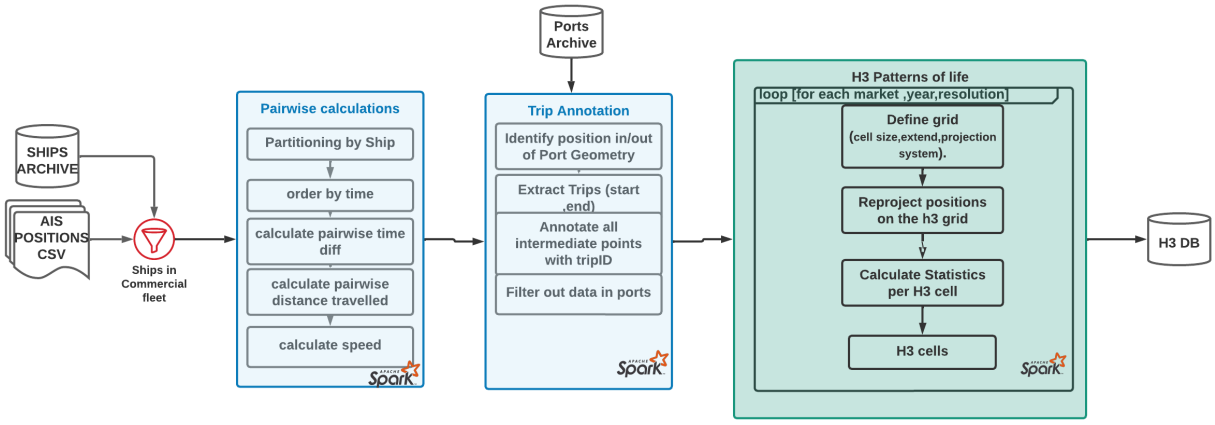

Figure 3: Execution flow diagram for the calculation of patterns of life.

and comparable results for cells in proximity. Even though an equal-area grid would be more appropriate for comparisons between distant cells, this is not a hard requirement. The AIS reception is affected by numerous factors. Thus, the calculated statistical representation of each cell corresponds to the observed feature distributions, so, as soon as cells in proximity have a fixed size we consider that any minor change of the cell size does not affect the accuracy of our methodology. If needed, explicit conversion using each cells area can be performed. In addition, we require the indexing mechanism to be performant and interoperable. These last two characteristics are of great importance for building extremely-scalable and efficient systems.

In this work we select the *H3* spatial index to address all the requirements for our analysis and future use of its results. The *H3* system is a GGDS system [19]. It relies on a hexagonal hierarchical grid and it provides a performant application interface implemented in a plethora of modern programming languages. The choice of hexagonal grids is advantageous for neighborhood analysis at scale. The neighborhood for *H3* corresponds to six adjacent neighbours at a fixed distance for each cell simplifying calculations against triangular and square grids as such spatial indexing systems have more neighbours and multiple distances per cell than hexagonal grid based indexing systems.

3.2.2 Execution Framework. We rely on the Apache Spark Framework (v3.1.2) to analyse the global AIS dataset we described in section 3.1. The execution framework setup consists of 128 v-cores capabilities, 256GB of RAM, a 1.7TB NVME disk that was used for cache purposes, 10TB of additional storage running Ubuntu OS[5].

3.3 Data Processing

In this section we present the key steps of data manipulation. For a pictorial representation of the process we refer the reader to Figure 2. A more detailed representation of the steps is shown in the execution flow diagram in Figure 3.

3.3.1 Data Cleaning and Preprocessing. The execution process starts with a preprocessing step that ensures the data quality and filters out irrelevant data entries. Then, we transform the dataset into vessel type aware entries. As soon as the position dataset is

[5]https://ubuntu.com/

Table 2: Grouping set (GS)

Group Identifier (GI)	Description
(H3-index)	Captures all traffic statistics crossing each cell.
(H3-index, vessel-type)	Statistical summary is broken down per cell and vessel type.
(H3-index, origin, destination, vessel-type)	Statistical summary is broken down per cell, origin, destination and vessel type.

loaded, it is initially partitioned based on the vessel identifier and then inspected to identify any values that are not within the expected by the protocol range. Values of *longitude*, *latitude*, *speed*, *course*, *heading* or *status* that do not comply with its expected value range are filtered out. To ensure that there are not out of order messages, we sort messages with respect to the reported timestamp for each vessel's partition. In addition, we calculate for consecutive messages of the same vessel their pairwise difference in time and their haversine distance. Based on those metrics we calculate each vessels *speed*, and filter out cases of non-feasible transitions (i.e transitions that require the vessel speed to exceed 50knots). In addition, we annotate all relevant rows with vessel static information, e.g the type, so that we can filter out messages from vessels that are not part of the commercial fleet or vessels that are outside of the target area (see Figure 2.a) and thus reduce the dataset size significantly. Figure 2 presents the implementation of the data cleaning and preprocessing steps.

3.3.2 Trip Semantics Extraction. Trip semantics are of great importance for the patterns of life methodology. We consider all messages of a specific vessel that have been captured in-between of consecutive two port stops to be part of the same trip. We rely on an external database to acquire port locations. Then, we perform a spatial technique, called geo-fencing in order to identify AIS records that are located within any port area. The first and the last records outside port-geometries are considered as the origin and destination timestamp respectively. We annotate all records with their trip identifier. Each trip, apart from the trip identifier, consists of the origin and destination port identifiers as well us the respected timestamps. Any message that cannot be annotated with trip information is excluded (see Figure 2.b) from further analysis.

For records that have trips semantics, as depicted in Figure 2.c, we can easily enrich the dataset with additional trip related features, such as the *elapsed time from departure* and the *actual time of arrival* by subtracting each message's reported timestamp respectively.

3.3.3 Projection to spatial index. Up until this point, the utilized partitioning mechanism has been the vessel identifier. To aggregate data, based on their local characteristics, we rely on the *H3* spatial index. We assign on each record the corresponding *H3* index based on the record location. The methodology can be repeated for resolution of any size due to the hierarchical structure of *H3*, given that we have enough data for the summaries be built from. For the results presented in this work we have selected, the resolutions of level 6 and 7, which create hexagons covering five and thirty six squared kilometres respectively. As the hexagon size increases, so does the index compression. The resolution level is selected so that cells are large enough to capture enough AIS messages and preserve statistical significance of the summaries and at the same time they preserve the sense of locality.

3.3.4 Feature Extraction and Processing. In this section we describe in detail the data structure and aggregation method. We also showcase that this specific data structure matches the capabilities of the selected execution framework. We first define the set of features that we group data upon. We refer to this feature set, as the grouping set (GS). All different combinations of values within the GS concatenated form the group identifier (GI). Some examples of selected identifiers are presented in Table 2. All records that belong to the same group are aggregated and assigned the same GI. We refer to the set of features that statistical summaries will be calculated on, as feature set (FS). The organisation of the features into groups and the calculation of summary statistics fits the MapReduce programming model, originally presented in [2], which is integral to the selected execution framework. The GS set corresponds to the mapping phase while the aggregated statistics correspond to the reduce phase.

In our case, the selected FS consists of three classes. The first class refers to the features that are directly extracted from AIS records. This class includes the number of records, the vessel identifiers, the reported course, speed and heading. The second class originates from trip semantics, described earlier in Section 3.3.2, it includes the trip identifier, the elapsed time from origin (ETO), the actual time to arrival (ATA), the port of origin and the port of destination. The last class refers to the transitions which is a summation of individual transitions from a cell to another with respect to the original order of AIS messages within each trip.

Table 3: Feature set (FS) and statistics

.

Description	Cnt	Dist	Mean	Std	Perc.	Bins	Top-N
Records	X						
Ships		X					
Course			X*			X	
Heading			X*			X	
Speed			X	X	X		
Trips		X					
ETO			X	X	X		
ATA			X	X	X		
Origin							X
Destination							X
Transitions							X

In Table 3 we present a detailed mapping of the features and the statistics that serve the needs of some common maritime use cases,

visited later in the results section. *Cnt*, *Dist*, *Mean* and *Std* stand for count, distinct count, mean and standard deviation statistics. We denote as *Percentiles* the 10^{th}, 50^{th} and 90^{th} approximate percentiles of the corresponding feature distribution. With *Bins* we denote the counters of messages that are organized in fixed value ranges, namely the bins. In this work we use the *bins* to split heading and course into 30 degrees counters. In addition, the *Top-N* statistic captures and counts the *N* most frequent values.

4 RESULTS

4.0.1 Compression and global coverage. Table 1 showcases that the original 2022 commercial fleet dataset consist of approximately 2.7 billion records. With respect to the selected resolution level, after applying our methodology, it can be represented by 7.3 million and 42.47 million cells for the 6^{th} and 7^{th} resolution levels of *H3* respectively. In terms of records, a full table scan would be needed for the online calculation of the complete feature set in Table. 3 for a specific location. The existence of the global inventory corresponds for the 6^{th} and 7^{th} resolution levels to 99.7% an 98.4% less hits respectively. Another important finding is that the commercial vessels of the dataset covered 51.69% of the total h3 cells available for the 6^{th} h3 resolution level and approximately 10% less if the child resolution h3 level is selected (i.e. the 7^{th}). This behaviour indicates that as the cell size of our analysis decreases, gaps might appear in the inventory.

Table 4: Coverage and Compression results for 2022 commercial fleet AIS dataset.

H3 resolution	#Cells	Compression	H3 Utilization
6	7.3 million	99.73%	51.69%
7	42.47million	98.44%	42.96%

4.1 Use cases

We showcase the usability and added value of the presented method through three different applications. First, we present how selected features of the resulting inventory can be used for local and global knowledge extraction. Then, we focus on how ATA and ETO present a baseline statistic for estimation of arrival time (ETA) for known sea routes. Additionally, we explicitly create an approach for route forecasting given that the origin and destination ports are known.

4.1.1 Knowledge extraction. Earlier in introduction section we presented the importance of understanding hidden patterns within AIS data. In this section we focus on global and local representations of the results of our methodology. We demonstrate how a selection of features from inventory is used for pictorial representation of the AIS patterns in global (Figure 1) and local (Figure 4) scope. In Figure 4, we use color mapping to visualise the average speed, average course and trip frequency. Anticipating the results for the Baltic sea area for all commercial traffic, we note that a proper filtering process reveals domain specific patterns representing the routes from the trip frequency (top), the loitering areas from the speed (middle) and the traffic separation schema from the course (bottom). Those patterns dynamically change over time and they are crucial for the understanding of the logistic chain efficiency. Apart from the pictorial representation, the statistical summaries are available for each cell and combination of the GI enabling more advanced calculations.

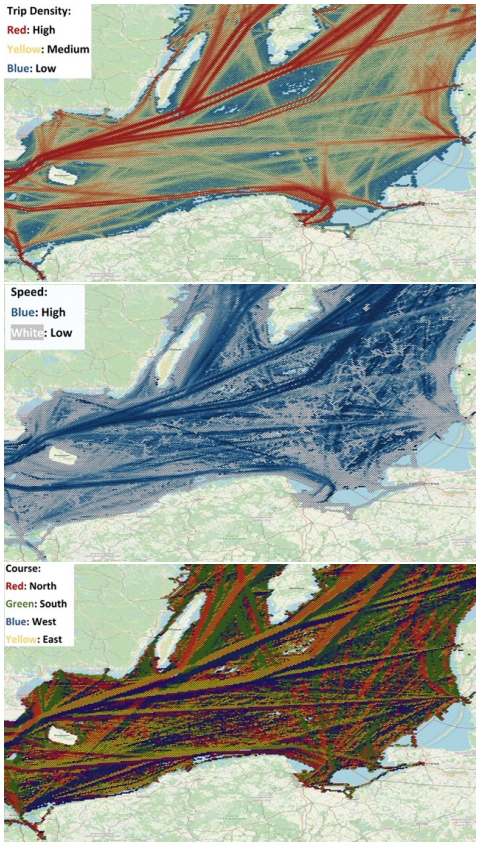

Figure 4: Visualisations of patterns of life for an area in the Baltic sea for 2022.

4.1.2 Estimated Time of Arrival. A long term challenge in the maritime analytics is the estimation of the time of arrival (ETA). This is a challenging problem in maritime world with multiple commercial and financial factors affecting the estimation performance. The utilization of historical AIS information is common practice in the domain, however, to the best of the authors knowledge, there is no previously published work of a global scale inventory that relies on the ATA of historical trips to estimate the expected time to destination. In Figure 5 we present in global scale, the average time to destination for each cell in *H3* resolution of level 6. We note that explicit statistics for ATA and ETO are also available for all value combinations of GI on each cell for online querying, each result set can be considered as a basic ETA estimate and they can be provided as input to more advanced ETA estimators.

4.1.3 Destination Prediction and Route Forecasting. Destination prediction and route forecasting are common challenges that both depend one on the other. We show that the global inventory build with this method is relevant to both challenges, touching only the

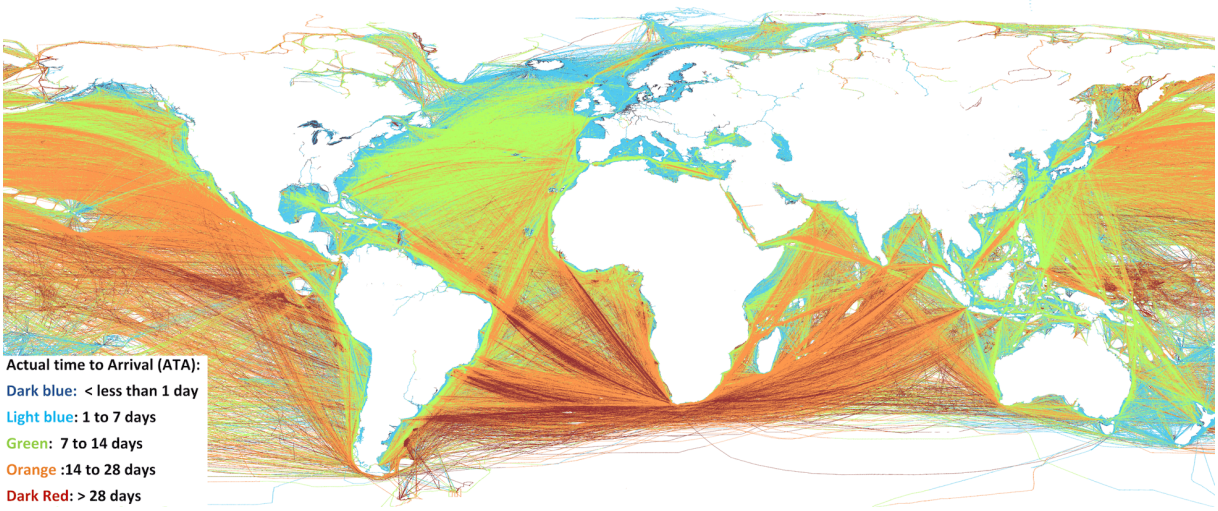

Figure 5: Patterns of Life, color depicts the average actual time to destination for all vessels per cell (H3 resolution 6).

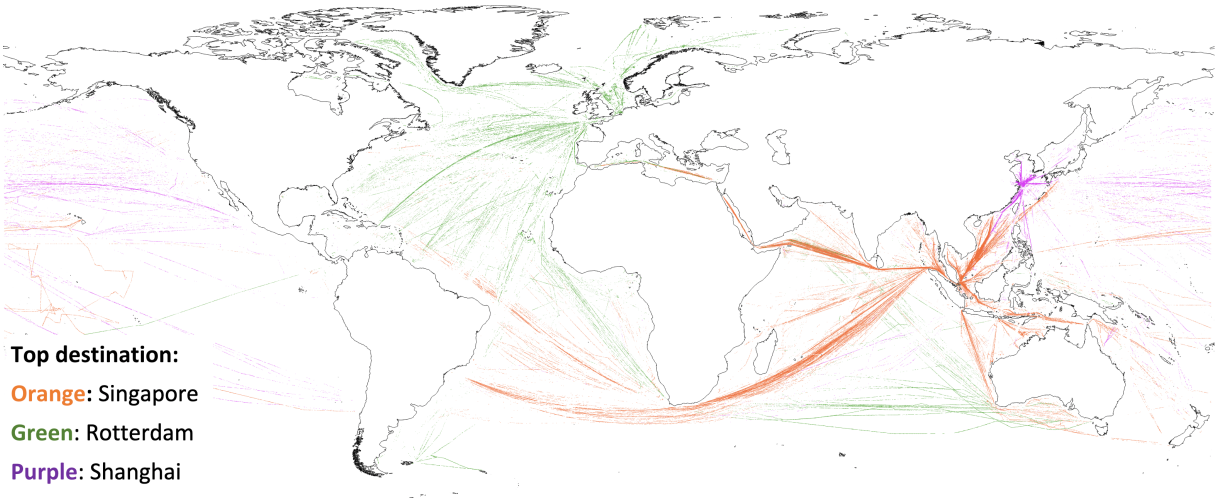

Figure 6: Patterns of Life cells where the most frequent destination for 2022 was Singapore in dark orange, Shanghai in purple and Rotterdam ports in green.

surface of them in this work. In Figure 6 we present a filtered version of the inventory depicting cells where to the most frequent destination is either the port of Singapore, the port of Shanghai or the port of Rotterdam. We note that the cell distribution is sparse, however the routes vessels follow towards those ports for 2022 are evident.

The routes boundaries and the underlying data on each cell can be compared against live AIS data for vessels travelling to those destinations. Given a stream of AIS positional reports of a vessel that her crew has not disclosed its destination, a streaming application may query online the inventory for each AIS message and retrieve the top-N destinations for vessels of the same type that sailed nearby in the past. In addition, it can keep track of this list, as the stream of AIS messages proceeds, to decide on the most probable destination.

Similarly for the route forecasting, given a vessel performing a specific origin-destination trip and her latest AIS positional report, we define an key, that consists of origin, destination and vessel-type. We query the global inventory to retrieve the full set of cells for which the key exists. The result set of cells corresponds to the full set of possible transition locations for the selected key, and it can be organized in a graph online, likewise Figure 2.f. For the graph representation, the vertices correspond to cells identifiers (i.e *H3* indices) and their connections are defined with respect to the transitions feature in Table 3. Given the graph, typical graph theory solutions that address the shortest path problem ,such as the *A-star* [6], can be used to forecast the route.

5 CONCLUSION AND FUTURE WORK

In this work, we present a multi-step methodology that relies on the *H3* spatial index and on the *Apache Spark* big data analytics engine to create a global scale inventory of statistical summaries from AIS data. We have applied out methodology on a 2022 global dataset consisting of 2.7 billion AIS messages originating from approximately 60 thousand commercial vessels. In this context, we presented pictorial representations of selected features, such as the average speed, course and heading, both in global and local scale, that can be used for extracting additional knowledge. In addition, we measured more than 98% of index compression, with the use of the spatial indexing, that allows end users to efficiently query the inventory. Apart from knowledge extraction capabilities of the generated inventory, we explored the possible utilization on time of arrival estimation, destination prediction and route forecasting challenges.

In future work, we intend to extend the proposed methodology to include features of non-AIS data. In this context, we plan to combine AIS with weather and commodity data in order to provide trade specific related summaries. Additionally, we aim to further explore hierarchical capabilities of the selected spatial index (H3) to provide non-uniform inventories but rather automatically adjusting to the density of maritime traffic, i.e., using larger cells in open sea areas which are known to have low vessel traffic density, preserving at the same time high resolution in dense areas, such as the ones near the ports.

ACKNOWLEDGMENTS

This research is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 101092749, project Critical Action Planning over Extreme-Scale Data (CREXDATA). This work has been developed on the basis of a global historical AIS dataset owned by Kpler. All software components that have been developed and presented within this work are owned by Kpler.

REFERENCES

[1] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. OPTICS: ordering points to identify the clustering structure. *ACM Sigmod record* 28, 2 (1999), 49–60.

[2] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: Simplified data processing on large clusters. *Commun. ACM* 51, 1 (2008), 107 – 113. Cited by: 11383; All Open Access, Bronze Open Access.

[3] Lubna Eljabu, Mohammad Etemad, and Stan Matwin. 2022. Spatial Clustering Method of Historical AIS Data for Maritime Traffic Routes Extraction. In *2022 IEEE International Conference on Big Data (Big Data)*. 893–902.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*, Vol. 96. 226–231.

[5] Dabo Guan, Daoping Wang, Stephane Hallegatte, Steven J. Davis, Jingwen Huo, Shuping Li, Yangchun Bai, Tianyang Lei, Qianyu Xue, D'Maris Coffman, Danyang Cheng, Peipei Chen, Xi Liang, Bing Xu, Xiaosheng Lu, Shouyang Wang, Klaus Hubacek, and Peng Gong. 2020. Global supply-chain effects of COVID-19 control measures. *Nature Human Behaviour* 4, 6 (June 2020), 577–587 (6).

[6] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics* 4, 2 (1968), 100–107.

[7] International Convention for the Safety of Life at Sea. [n. d.]. Solas Chapter V - Regulation 19 - Carriage requirements for shipborne navigational systems and equipment. https://assets.publishing.service.gov.uk/media/5a7f0081ed915d74e33f3c6e/solas_v_on_safety_of_navigation.pdf. [Accessed: 15-Nov-2023].

[8] R. Praveen Jain, Edmund F. Brekke, and Adil Rasheed. 2022. Unsupervised Clustering of Marine Vessel Trajectories in Historical AIS Database. In *2022 25th International Conference on Information Fusion (FUSION)*. 1–6.

[9] Kpler. 2023. Unlock Maritime Insights with MarineTraffic. https://www.kpler.com/product/maritime

[10] Jeong-Seok Lee, Woo-Ju Son, Hyeong-Tak Lee, and Ik-Soon Cho. 2020. Verification of Novel Maritime Route Extraction Using Kernel Density Estimation Analysis with Automatic Identification System Data. *Journal of Marine Science and Engineering* 8, 5 (2020).

[11] Huanhuan Li, Jingxian Liu, Ryan Wen Liu, Naixue Xiong, Kefeng Wu, and Taihoon Kim. 2017. A Dimensionality Reduction-Based Multi-Step Clustering Method for Robust Vessel Trajectory Analysis. *Sensors* 17, 8 (2017).

[12] Huanhuan Li, Jingxian Liu, Kefeng Wu, Zaili Yang, Ryan Wen Liu, and Naixue Xiong. 2018. Spatio-Temporal Vessel Trajectory Clustering Based on Data Mapping and Density. *IEEE Access* 6 (2018), 58939–58954.

[13] Antonios Makris, Konstantinos Tserpes, Giannis Spiliopoulos, Dimitrios Zissis, and Dimosthenis Anagnostopoulos. 2021. MongoDB Vs PostgreSQL: A comparative study on performance aspects. *GeoInformatica* 25 (2021), 243–268.

[14] L. Moriarty. 2020. Public health responses to COVID-19 outbreaks on cruise ships — Worldwide, February–March 2020. *Morbidity and Mortality Weekly Report (MMWR)* 69 (2020), 347–352.

[15] Giuliana Pallotta, Michele Vespe, and Karna Bryan. 2013. Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction. *Entropy* 15, 6 (2013), 2218–2245.

[16] Cyril RAY, Richard DRÉO, Elena CAMOSSI, and Anne-Laure JOUSSELME. 2018. *Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance.* https://doi.org/10.5281/zenodo.1167595

[17] Vasja Roblek, Maja Meško, and Alojz Krapež. 2016. A Complex View of Industry 4.0. *SAGE Open* 6, 2 (2016), 2158244016653987. https://doi.org/10.1177/2158244016653987

[18] H. Rong, A.P. Teixeira, and C. Guedes Soares. 2022. Maritime traffic probabilistic prediction based on ship motion pattern extraction. *Reliability Engineering & System Safety* 217 (2022), 108061.

[19] Kevin Sahr, Denis White, and Jon A. Kimerling. 2003. Geodesic Discrete Global Grid Systems. *Cartography and Geographic Information Science* 30 (2003), 121 – 134.

[20] Giannis Spiliopoulos, Konstantinos Chatzikokolakis, Dimitrios Zissis, Evmorfia Biliri, Dimitrios Papaspyros, Giannis Tsapelas, and Spyros Mouzakitis. 2017. Knowledge extraction from maritime spatiotemporal data: An evaluation of clustering algorithms on Big Data. In *2017 IEEE International Conference on Big Data (Big Data)*. 1682–1687.

[21] Andreas Tritsarolis, Yannis Kontoulis, and Yannis Theodoridis. 2022. *The Piraeus AIS Dataset for Large-scale Maritime Data Analytics.* https://doi.org/10.5281/zenodo.6323416

[22] United Nations Conference on Trade and Development (UNCTAD). 2017. Review of Maritime Transport. https://unctad.org/en/pages/PublicationWebflyer.aspx?publicationid=1890.

[23] United Nations Conference on Trade and Development (UNCTAD). 2020. COVID-19 and maritime transport: Impact and responses. https://unctad.org/system/files/official-document/dtltlbinf2020d1_en.pdf.

[24] United Nations Conference on Trade and Development (UNCTAD). 2021. Review of Maritime Transport. https://unctad.org/publication/review-maritime-transport-2021.

[25] United Nations Conference on Trade and Development (UNCTAD). 2022. Review of Maritime Transport. https://unctad.org/system/files/official-document/rmt2022_en.pdf.

[26] United Nations Conference on Trade and Development (UNCTAD). 2023. Review of Maritime Transport. https://unctad.org/publication/review-maritime-transport-2023.

[27] Yuanqiao Wen, Zhongyi Sui, Chunhui Zhou, Changshi Xiao, Qianqian Chen, Dong Han, and Yimeng Zhang. 2020. Automatic ship route design between two ports: A data-driven method. *Applied Ocean Research* 96 (2020), 102049.

[28] Wenjing Yan, Rong Wen, Allan N. Zhang, and Dazhi Yang. 2016. Vessel movement analysis and pattern discovery using density-based clustering approach. In *2016 IEEE International Conference on Big Data (Big Data)*. 3798–3806.

[29] Zhaojin Yan, Yijia Xiao, Liang Cheng, Rong He, Xiaoguang Ruan, Xiao Zhou, Manchun Li, and Ran Bin. 2020. Exploring AIS data for intelligent maritime routes extraction. *Applied Ocean Research* 101 (2020), 102271.

[30] Yuan-qiang Zhang and Guo-you Shi. 2021. Trajectory Similarity Measure Design for Ship Trajectory Clustering. In *2021 IEEE 6th International Conference on Big Data Analytics (ICBDA)*. 181–187.

[31] Yu Zheng. 2015. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6, 3 (2015), 1–41.

[32] D. Zissis, K. Chatzikokolakis, G. Spiliopoulos, and M. Vodas. 2020. A Distributed Spatial Method for Modeling Maritime Routes. *IEEE Access* 8 (2020), 47556–47568.