# Scalable and distributed sea port operational areas estimation from AIS data

Leonardo M. Millefiori*, Dimitrios Zissis*†, Luca Cazzanti* and Gianfranco Arcieri*
*NATO STO Centre for Maritime Research and Experimentation (CMRE), La Spezia, Italy
Email: {leonardo.millefiori, dimitrios.zissis, luca.cazzanti, gianfranco.arcieri}@cmre.nato.int
†Department of Product and Systems Design Engineering, University of the Aegean, Syros, Greece
Email: dzissis@aegean.gr

***Abstract*—Seaports are spatial units that do not remain static over time. They are constantly in flux, evolving according to environmental and connectivity patterns both in size and operational capacity. As such any valid decision making regarding port investment and policy making, essentially needs to take into account port evolution over time and space; thus, accurately defining a seaport's exact location, operational boundaries, capacity, connectivity indicators, environmental impact and overall throughput. In this work, we apply a data driven approach to defining a seaport's extended area of operation based on data collected though the Automatic Identification System (AIS). Specifically, we present our adaptation of the well-known KDE algorithm to the MapReduce paradigm, and report results on the port of Rotterdam.**

***Keywords*-big data, KDE, AIS, port location estimation, Rotterdam port, Apache Spark, MapReduce**

## I. Introduction

The gateway to European commerce and industry is one of the most complex and dense port networks in the world. These hubs connect seaborne trade and passengers to the multimodal transportation networks in Europe, which include rail, road, air and inland waterways. A recent report published by the European Commission calculated that 74% of goods imported and exported and 37% of exchanges within the Union transit through seaports [1], [2]. Interestingly though, from the roughly 1200 ports along the 70 000 km EU coastline, approximately 20% of this traffic is served by only three ports: Rotterdam, Antwerp and Hamburg [3]. Today EU ports are facing a number of unique challenges, which include [1], [2]

- A significant increase in vessels size, speed and volume; in particular ultra-large container ships, new types of Roll-on/Roll-off (RORO) ferries and gas-carriers, are putting pressure onto traditional port structural infrastructures, while increasing maritime risks, as these vessels are difficult to maneuver in small shipping lanes;
- Significant developments in the energy trades, with a shift from oil and refined products towards gas that require significant gasification facilities in ports and potential volumes of dry biomass and CO2 transport and storage;
- Adherence to stricter requirements on environmental performance and alternative fuels.

In the midst of an economic recession, significant investment is required to keep the EU ports competitive, including extensions of berths, deepening of basins and canals to enable large vessel maneuvering, new terminal and operational procedures to allow for parallel coordination of services, implementation of Information and communications technology (ICT) logistic systems and supply management and much more. According to the EU commission, the complexity of the administrative procedures for customs clearance results in big delays at ports, representing just another major obstacle to the competitiveness of sea shipping and the efficiency of Union ports [4]. Additionally, port security has come into the public spotlight, as ports are critically important infrastructures and potential gateways for unlawful trades concerning drugs, weapons, counterfeited goods and even explosive materials.

Recognizing the vital role the ports play in the Union's economy, international competitiveness and the single market integration, a number of EU wide policies and directives are being implemented, which aim at guaranteeing the availability of a well-connected port infrastructure. As such, the European Transport Infrastructure Plan identifies 329 ports of common interest, including 104 ports of strategic interest, 9 multimodal core network corridors that start and end in seaports and reserves a budget of 26 billion EUR for the period 2014–2020. It is the objective of the European port policy, while respecting the overarching need for better infrastructure planning, to stimulate competition among ports and ensure that adequate capacity is available for a sustainable growth of European trade and effective port operations [3]. It is important to note though that although investments are necessary, simultaneously the World Economic Forum states that, in transport, only 40% of load capacity is used nowadays. Now more than ever, it is a necessity to examine the potential of enhancing the efficiency of the logistic chain by exchanging data, by

This work is supported by the NATO Supreme Allied Command Transformation (SACT) under project SAC000608 – Data Knowledge Operational Effectiveness, by Microsoft Research through a Microsoft Azure for Research Award, and by MarineTraffic.

exploring and exploiting the benefits of digitalization [5]. Increasing the capacity and redesigning ports, at the cost of the taxpayer, should be contained at no more than what is operationally necessary [6]. Similar to other industries, decisions regarding the redesign of ports areas, their increase in operational capacity and infrastructure, need to be based on measurable data which can be transformed into actionable information.

Data science involves principles, processes, and techniques for understanding phenomena with the ultimate goal of improving decision making, as this generally is of paramount interest to business. Data-driven decision-making (DDDM) refers to the practice of basing decisions on the analysis of data rather than purely on intuition [7]. While in the past, sea transport surveillance had suffered from a lack of data, current tracking technology has transformed the problem into one of an overabundance of information, leading to a need for automated analysis. The major challenge faced today is developing the ability to identify patterns emerging within these huge datasets, fused from a variety of sources and generated from monitoring a large number of vessels on a day-to-day basis. The extraction of implicit and often unknown information from these datasets belongs to the field of data mining and data science. Progressively huge amounts of structured and unstructured data, tracking vessels during their voyages across the seas are becoming available, mostly due to the Automatic Identification System (AIS) that vessels of specific categories are required to carry. These datasets provide detailed insights into the patterns that vessels follow, while they can operate as benchmarking tools for port authorities regarding the effectiveness and efficiency of their ports. A data-driven approach to seaborne transportation could potentially,

- Accomplish a decrease in port congestion and seaways by monitoring and improving the forecasting of vessel arrivals (ship sizes, cargoes, ETAs, loading/discharge times) to enable better planning and execution of port operations (virtual arrivals);
- Achieve a reduction of accidents at sea by timely detecting hazardous situations (including malicious events) based on the vessels trajectory, behavior (e.g. speed, course) and context (e.g. weather conditions or nearby ships), while proposing measures of proactive prevention; and
- Contribute to the reduction of green-house gas (GHG) emission intensity for each vessel by providing data to increase and optimize operational efficiency.

Benchmarking ports will support greater transparency in policy making, stakeholder decision making, public funding while promoting healthy competition between the ports themselves. Benchmarking measurements include maritime connectivity indicators, current port operational capacity, number of port calls, type of vessels, call size, cargo throughput, intermodal connectivity, vessel time at anchorage outside port, number of vessel waiting to be processed, while taking of course into consideration the specific characteristics of regions and other port externalities. Generating valid and reliable measurements though regarding ports statistics is a highly complex task. We often overlook the fact that maritime networks operate as "small worlds", where content and size varies over space and time, under the influence of the trade and carrier patterns. Such spatial units are often not well defined and delimited, such as port region, port system, port range as they evolve according to patterns [8]. The stepping stone for any useful analytics and data driven approach to port planning is accurately defining a seaport's location and operational boundaries, so capacity and efficiency can be calculated. This paper discusses work in progress at the NATO STO Centre for Maritime Research and Experimentation (CMRE) to estimate port areas in a scalable, data-driven way. Knowing the extent of port areas is an important component of larger maritime traffic analysis systems that employ computational techniques to achieve Maritime Situational Awareness (MSA). For example, accurately detecting which vessels visit a given port and its surrounding areas enables the study of vessel traffic Patterns of Life (PoL's) in a region, the calculation of summary statistics on the volume and type of vessels, and the detection of discrepancies in the vessel-declared origin and destination ports.

In our approach, we exploit the large volume of historical and real-time AIS data to estimate the port areas in a data-driven way, with minimal reliance on other sources of information. However, as the amount of available AIS data grows to massive scales, computational techniques for MSA —which we call computational MSA— must also contend with acquiring, storing, and processing the data. We are addressing these challenges by leveraging a cluster of computers to store the AIS data and to serve the spatial clustering and density estimation operations underlying the proposed port area algorithm. The proposed approach can be extended to other types of areas besides ports: offshore platforms, anchorage areas, and fishing grounds can be detected automatically and their extent estimated using this approach. This is particularly beneficial because often these types of areas have dynamic boundaries that change with the seasons, as a consequence of newly introduced local vessel traffic schemes, or as new maritime support facilities become available. Thus, being able to estimate automatically and quickly the current extent of stationary areas worldwide becomes essential.

## II. Background and related work

In recent years extensive research has been performed in exploring methods of increasing the effectiveness and efficiency of ports through ICT; numerous research efforts have focused around the themes of ICT as a method of supporting

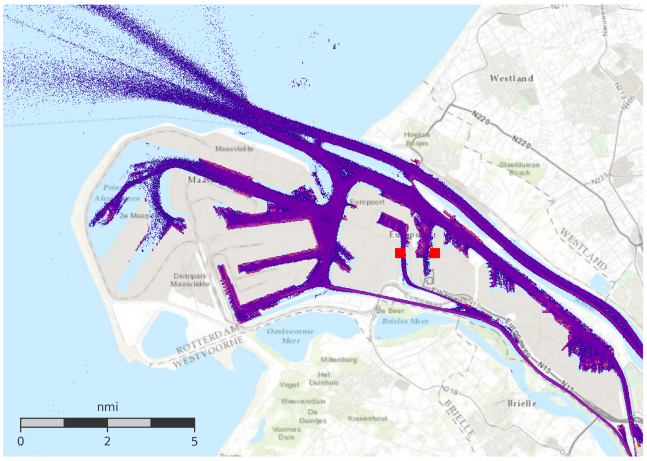



Figure 1. Density of AIS messages collected by MarineTraffic from January to March 2015. Each pixel covers a 10-by-10 m square on the ground and its color is (logarithmically) proportional to the number of AIS messages whose reported positions fall within its footprint.

more efficient logistics supply chain management, security management, greener performance and port benchmarking.

At an EU level, a number of collaboration research projects have been funded related to improving the efficiency and effectiveness of sea operations either through the FP6, FP7 or Horizon 2020 instruments. As such the SAIL project, which was funded under FP7, designed and developed a novel system aimed at improving integrated logistics management and decision support for intermodal port and dry port facilities. The two year PPRISM project, co-funded by the European Commission, delivered a shortlist of indicators that form the basis of a future European Port Observatory, which will take the form of a Port Sector Performance Dashboard. This work was later followed up by the PORTOPIA (Ports Observatory for Performance Indicator Analysis) project, funded also under FP7, whose main objective was to develop an enhanced ports observatory with a set of indicators measuring EU ports performance, activities and developments. The SUPPORT project aimed at providing general methods, technology and training services to be used by any European Port to upgrade their security capability. In relation to security, various research projects have been funded, such as the Maritime Navigation and Information Services (MarNIS), Motorways & Electronic Navigation by Intelligence at Sea (MONALISA), Advanced National Networks for Administrations (ANNA), Vessel traffic monitoring in EU waters (SafeSeaNet) and MOS.

AIS data has proven to be a valid method for monitoring vessels and extracting valuable information regarding vessel behavior, operational patterns and performance statistics. As Tichavska, Cabrera, Tovar and Arana point out, AIS data in research has been used for a variety of applications including optimization of radio propagation channel techniques, real-time statistical processing of traffic information, improving ship traffic management and operations, sustainable transport solutions and many more [9]. Pallotta, Vespe and Bryan make use of AIS data for vessel pattern knowledge discovery as a framework for anomaly detection and route prediction [10]. Huijbrechts, Velikova, Michels and Scheepens present their work performed in the scope of the METIS project that makes use of AIS data for automated and consolidated "situational understanding" [11]. Rashidi and Koto predict the CO2 emitted by marine transport in Batam-Singapore Channel using AIS data [12]. In [13], Ristic, La Scala, Morelande and Gordon, use AIS data to extract motion patterns which are then used to construct the corresponding motion anomaly detectors. In relation to sea ports research and AIS, Ricci, Marinacci and Rizzetto, made use of AIS to model maritime terminals operations, specifically focusing on the port of Messina [14]. While Donateo, Gregoris, Gambaro, Merico, Giua, Nocioni and Contini evaluated the contribution of harbor activities and ship traffic to PM2.5, particle number concentrations and PAHs in a port city of the Mediterranean Sea (Italy) [15]. Jing and Shuang perform a safety evaluation of China's maritime transport key nodes based on AIS [16].

However, to the best of our knowledge, much less work has been performed in relation to using AIS data to define the exact seaport location and its operational boundaries, so capacity and efficiency can be calculated. In this paper we propose a method to estimate port locations and operational areas in a scalable and unsupervised way, using the Kernel Density Estimator (KDE) and taking advantage of one of the most widely adopted distributed programming models.

In this work, our focus is on the port of Rotterdam. Located in South Holland, within the Rhine-Meuse-Scheldt river delta at the North Sea and provides high-frequency connections to numerous destinations across Europe. The port's annual throughput amounts to some 465 million tonnes, while the port area is more than $40\,\mathrm{km}$, making it the largest port in Europe and ninth in the world. From the port of Rotterdam an extensive fleet of inland vessels transports cargo via the Maas and the Rhine directly to the major economic centers in the Netherlands, Germany, Belgium, France, Switzerland and Austria. On a yearly basis more than $30\,000$ seagoing vessels and $110\,000$ inland vessels visit the port. It is also one of the most complex ports in the world with numerous terminals including 6 deep sea container terminals, 3 short-sea terminals and 18 empty depots; 6 RORO and 19 general cargo terminals, 17 dry bulk terminals, numerous oil refineries, chemical locations, gas and power terminals and other.

## III. Approach

### A. Data description

The AIS was originally conceived as a navigational safety system to support vessel traffic services in ports and harbours, but soon after its adoption, and especially after the

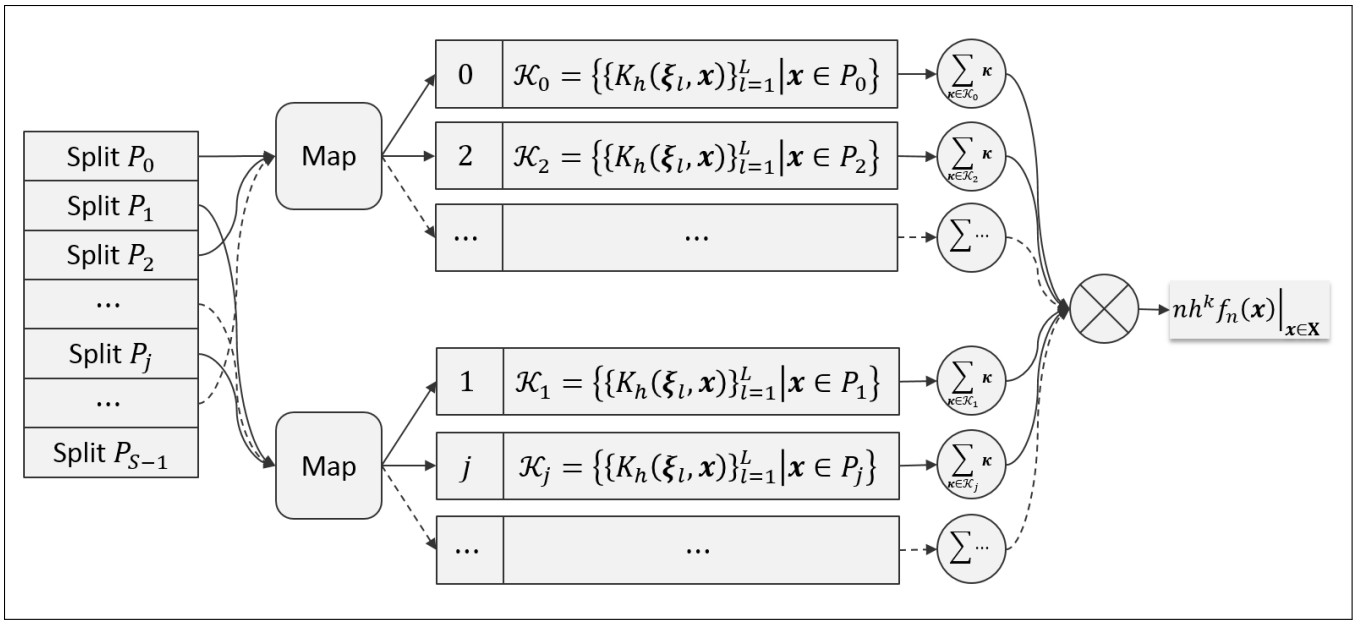



Figure 2. Implementation of the KDE with MapReduce logic. The leftmost block shows the the input data relative to a single port, which is organized in partitions and stored on a DFS or a distributed database. Taking advantage of its linearity, the computation of the KDE can be distributed among multiple nodes, each of which performs an expansion of the input partition with the Gaussian kernel.

International Maritime Organization (IMO) mandated AIS transceivers to be installed onboard a significant number of commercial vessels [17], AIS began being used also to achieve broader MSA [18], which is the understanding of the factors that impact the economy, environment, security, and safety of the maritime domain.

The AIS communication protocol is asynchronous and prescribes that different types of messages be transmitted with different frequencies. In fact, the ITU 1371-4 standard defines 64 different types of AIS messages that can be broadcast by AIS transceivers. In this work we focus on the 6 most relevant ones for MSA, which account for approximately 90% of AIS typical scenarios [19]. Types 1, 2, 3, 18, and 19 are position reports, which include latitude, longitude, speed-over-ground (SOG), course-over-ground (COG), and other fields related to ship movement; type 5 messages contain static-and-voyage information, which includes the IMO identifier, radio call sign, name, ship dimensions, ship and cargo types. In all messages, each vessel is identified by its Marine Mobile Service Identifier (MMSI) number.

Since the first introduction of AIS, maritime traffic and global compliance with the international requirements have steadily increased, and the many worldwide networks of AIS coastal receivers have grown, resulting in larger and larger volumes of AIS data. Every month, this amounts to many millions of AIS messages produced by hundreds of thousands of unique vessels [20].

In the remainder of this paper, we use a dataset about 3.4 million AIS messages made available by MarineTraffic and recorded from January to March 2015 in an area of about 300 square km over the port of Rotterdam. In Fig. 1 we report, for reference, a density map of the dataset over the area of interest; each pixel in the figure covers a 10-by-10 m square on the ground and its color is proportional to the logarithm of the number of recorded AIS messages whose reported positions fall within its footprint.

### B. Kernel density estimation

Let us assume that $\boldsymbol{x}_i \in \mathbb{R}^k$, with $i = 1, \ldots, n$, are a set of observations from a probability density $f$. Initially introduced by Rosenblatt [21], a basic KDE of $f$ has the form [22]:

$$f_n(\boldsymbol{x}) = \frac{1}{nh^k} \sum_{i=1}^{n} K_h(\boldsymbol{x}, \boldsymbol{x}_i), \tag{1}$$

where $K_h$ is the kernel function, and $h$ denotes the kernel bandwidth (or window width), which is a smoothing parameter. The choice of $h$ has a strong influence on the estimate, because different values highlight different features of the data, depending on the density under consideration. The choice of a kernel function, on the other hand, is not crucial to the statistical performance, and a widely adopted choice is the Gaussian kernel, defined as below

$$K_h(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{(2\pi)^{\frac{k}{2}} \sqrt{|\boldsymbol{\Sigma}|}} \mathrm{e}^{-\frac{(\boldsymbol{p}-\boldsymbol{q})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{p}-\boldsymbol{q})}{2h^2}}. \tag{2}$$

*1) Convolution:* Apart from a scaling factor, the KDE formula (1) can also be seen as a convolution (which we denote with the $*$ operator) between the empirical Probability Density Function (PDF) and the kernel function [23],

that is

$$\begin{aligned}\phi_n * K_h &= \int_{\mathbf{D}} \left( \frac{1}{n} \sum_{i=1}^{n} \delta\left(\boldsymbol{\xi} - \boldsymbol{x}_i\right) \right) K_h\left(\boldsymbol{x} - \boldsymbol{\xi}\right) \, \mathrm{d}^k \boldsymbol{\xi} \\ &= \frac{1}{n} \sum_{i=1}^{n} K_h\left(\boldsymbol{x} - \boldsymbol{x_i}\right) = h^k f_n(\boldsymbol{x}),\end{aligned} \tag{3}$$

where $\phi_n$ is the empirical PDF, expressed as a sum of $n$ Dirac delta functions $\delta(\cdot)$ centered in the data samples. A computationally efficient variant of this formulation bins the data samples into $k$-dimensional histograms, and convolves the histogram with the kernels instead of the individual delta functions. This variant is appealing when the data size increases, because it produces an essentially identical result at a fraction of the computational cost.

*2) Adaptive KDE:* Both the KDE in (1) and the KDE by convolution (3) employ a fixed kernel bandwidth for all the observed data points. An intuitive improvement is to weight observations non uniformly; that is, extreme observations in the tails of the distribution should have their mass spread in a broader region than those in the body of the distribution. Specifically, instead of having a single value for $h$, in the adaptive KDE approach $h_i$, for $i = 1, \ldots, n$, is the bandwidth of the kernel centered in the $i$-th observation.

The first challenge is *how* to decide if an observation belongs to a region of high or low density. The adaptive approach [23] relies in fact on a two-stage procedure: combining (1) with (2), a pilot estimate is first computed to identify low-density regions coarsely, using a fixed bandwidth factor. Since only a coarse idea is required of how the density is distributed in the area of interest, here we can use the convolved histogram (3), which comes at a fraction of the computational cost required to compute (1).

*3) Local bandwidth factors:* Under the assumption that the underlying distribution is $k$-variate normal, the optimum (fixed) window can be written as [23]:

$$h^* = \left( \frac{4}{n(k+2)} \right)^{\frac{1}{k+4}}. \tag{4}$$

The *local bandwidth* factors $\lambda_i$, for $i = 1, \ldots, n$ are then given by

$$\lambda_i = \left( \frac{f_n\left(\boldsymbol{x}_i\right)}{g} \right)^{-\alpha}, \tag{5}$$

where $0 \leq \alpha \leq 1$ is the sensitivity parameter and $g$ is the geometric mean of the fixed-bandwidth density estimate $f_n\left(\boldsymbol{x}_i\right)$ evaluated in the data points

$$\log g = \frac{1}{n} \sum_{i=1}^{n} \log f_n\left(\boldsymbol{x}_i\right). \tag{6}$$

The adaptive KDE of $f$ can be finally expressed as

$$\hat{f}_n\left(\boldsymbol{x}\right) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\left(h^* \lambda_i\right)^k} K_{h^* \lambda_i}\left(\boldsymbol{x}, \boldsymbol{x}_i\right). \tag{7}$$

## IV. Implementation and results

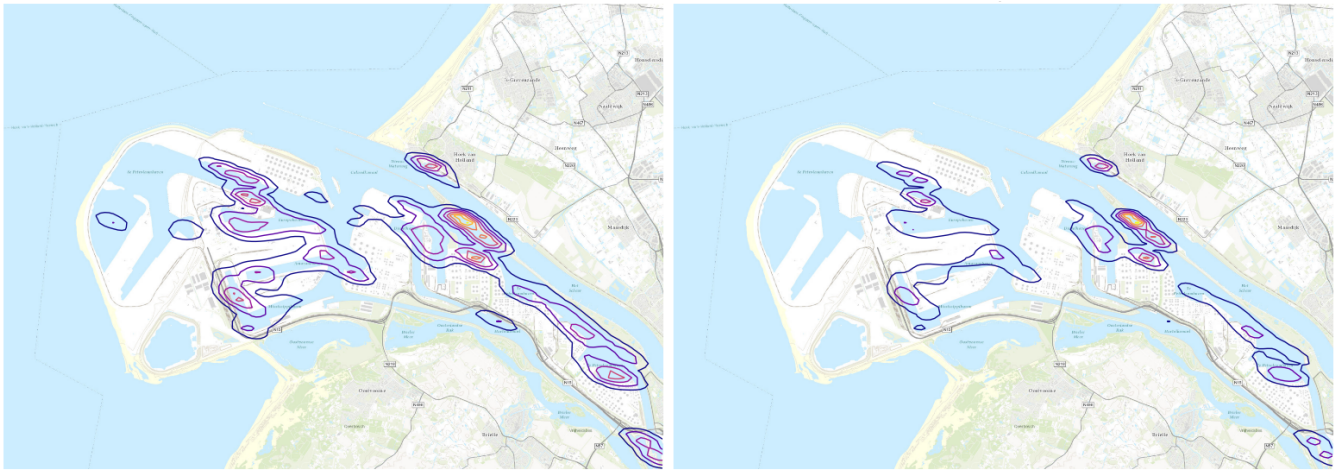
Let us indicate now the *full* kinematic state of a vessel at a generic time with $\boldsymbol{\chi}_i = [a_i,\, b_i,\, v_i]^{\mathsf{T}} \in \mathbb{R}^3$, where $a$ and $b$ represent the longitude and latitude coordinates, respectively, of the ship in a geographic coordinate system, and $v \geq 0$ is the instantaneous speed of the vessel. We introduce also a *reduced* vessel kinematic state that doesn't include the instantaneous speed $\boldsymbol{x}_i = [a_i,\, b_i]^{\mathsf{T}} \in \mathbb{R}^2$. Finally, we observe the ship traffic in the neighborhood of a port in the time interval $[0, T]$, where $T$ can be hours, days or even months, depending on the application.

Our objective is to determine the area of the port given the set of AIS observations $\mathcal{X} = \{\boldsymbol{\chi}_i\}_{i=1}^{n}$, that can be made up either by the full or reduced kinematic states of the ships observed in the area of interest. Assuming that the samples $\mathcal{X}$ are drawn from a probability density function $f$, the proposed approach consists of applying the KDE to the data samples, and determining the port extent using horizontal cuts of the resulting estimated probability density function.

Unfortunately, the direct computation of the fixed KDE (1) is highly inefficient, especially for large or highly dimensional data sets. In fact several approaches have been proposed in the past to reduce the computational burden [24]–[26]. However, as the data set size and its dimensionality increase, even the aforementioned approaches can easily become computationally prohibitive and therefore distributed approaches are necessary. Zheng et al. [27] have recently proposed randomized and deterministic distributed algorithms for efficient KDE with quality guarantees, adapting them to the popular MapReduce [28] programming model. As in [27], our approach is to take advantage of the linearity of the KDE to distribute the computation among many different nodes using the MapReduce distributed programming model.

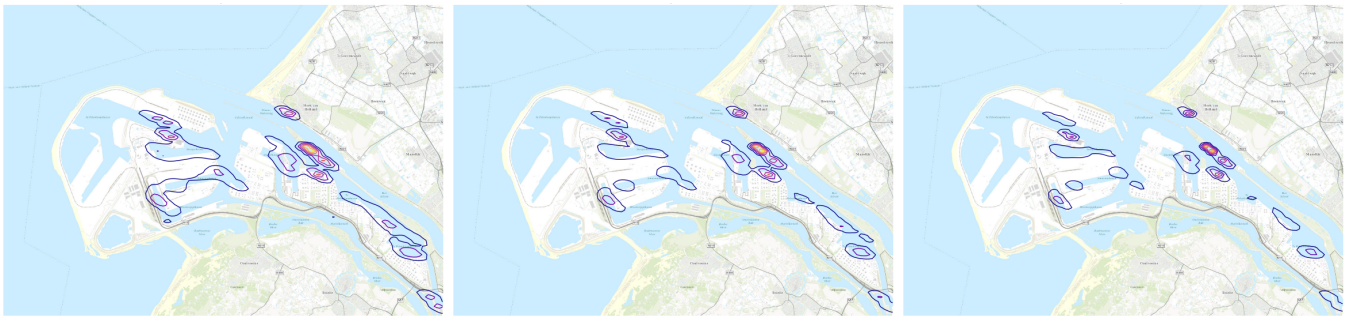
In Fig. 2 we report a conceptual representation of the formulation of kernel density estimation problem in the MapReduce framework. The leftmost blocks represent the partitions of the input data relative to a single port. The problem of associating each data sample with the corresponding port is a separate issue, that can be easily addressed using, for instance, a k-Nearest Neighbor (k-NN) classifier. Taking advantage of the linearity of the KDE, each Map function produces an expansion of the given input partition with the Gaussian kernel. Finally, the Reduce step is responsible for summing up all the contributes and eventually produces the final estimate. In the adaptive version, this schema is expanded with the computation of the local bandwidth factors, that are then associated to the corresponding data samples in the partitions.

For our purposes, we consider the port as the *extended* location where ships exhibit a very low speed. Consequently, there are two possible approaches for estimating the density function. One possibility is to compute the KDE in $\mathbb{R}^3$ at a



(a) Fixed-bandwidth KDE (b) Adaptive KDE

Figure 3. Comparison of fixed and adaptive bandwidth kernel density estimates computed in the Rotterdam port area. The fixed-bandwidth version (a) produces a smoother result, but is unable to deal satisfactory with the low-density regions. The adaptive KDE (b) has a higher computational cost than the fixed KDE, but it produces a *spikier*, and consequently narrower, estimate on low-density regions. Both the estimates have been computed on the available data collected by MarineTraffic from January to March 2015, having selected only those ships whose speed reported by the AIS was not exceeding the fixed threshold of $1\,\mathrm{kn}$.



(a) $\alpha = 0.3$ (b) $\alpha = 0.5$ (c) $\alpha = 0.7$

Figure 4. Effect of the sensitivity $\alpha$ on the resulting PDF estimated with the adaptive bandwidth approach. The three panels refer to three different sensitivity levels, namely $\alpha = 0.3$ (a), $\alpha = 0.5$ (b) and $\alpha = 0.7$ (c). Smaller values of the sensitivity parameter $\alpha$ produce results that are more similar to the fixed-bandwidth KDE, while *spikier* density functions are created with higher values of $\alpha$, but they are inevitable more fragmented.

very high computational cost using the complete kinematic states $\boldsymbol{\chi}_i$ including also the ship speed, and then compute the spatial density estimate $\bar{f}_n(\boldsymbol{x})$ by marginalization of $f_n(\boldsymbol{\chi})$

$$\bar{f}_n(\boldsymbol{x}) = \int_0^{v_T} f_n(\boldsymbol{\chi})\,\mathrm{d}v,$$

where $v_T$ is the speed threshold that discriminates the stationary ships from those under way. Unfortunately, this approach usually does not improve significantly the quality of the resulting estimate over less computationally demanding alternatives, especially in low-density regions.

The second possibility is to form the KDE in $\mathbb{R}^2$ using only the positional information $\boldsymbol{x}_i$ of the ships that can be considered stationary. In other words, given the set of all the observations, we can build a subset of the positional states of only those ships whose speed is below a desired threshold $v_T$, and compute the KDE on this subset. This approach can be also seen as an approximation to the first one that trades some result accuracy for a more affordable computational cost.

Filtering out all the ships whose velocity exceeds the threshold of $1\,\mathrm{kn}$ leaves us with a dataset of $\approx 1.5$ million samples, from an initial size of $\approx 3.4$ million. These $\approx 1.5$ million data samples are then used to compute the density estimate, corresponding to the AIS messages received by MarineTraffic from January to March 2015 whose reported velocity was below $1\,\mathrm{kn}$. Finally, we apply both the fixed and adaptive bandwidth KDE to this data set.

We rely on a Apache Spark™ cluster made up by: $11$ worker nodes, each one equipped with $4$ processing cores and $14\,\mathrm{GB}$ RAM; and $2$ head nodes, each one equipped with $8$ processing cores and $14\,\mathrm{GB}$ RAM, summing up to a

total of $60$ computing cores and $154\,\mathrm{GB}$ RAM. In our setup, the fixed KDE on the aforementioned area of interest takes about $3\,\mathrm{minutes}$. The adaptive KDE has as first step a fixed-bandwidth KDE and is more computationally expensive than the fixed KDE by definition, taking, with the aforementioned configuration, about $6\,\mathrm{minutes}$ to run.

Fig. 3 shows a comparison between the port area of Rotterdam computed with the fixed KDE (a) versus the one computed with the adaptive KDE (b). The estimates have been determined using the available data collected by MarineTraffic from January to March 2015 and the fixed-bandwidth approach in $\mathbb{R}^2$, having selected only those ships whose speed reported by the AIS was not exceeding the fixed threshold of $1\,\mathrm{kn}$. The horizontal cuts of the PDF surround the position of the port, as recorded in the World Port Index (WPI) [29].

As expected, the PDF exhibits multiple peaks, which are presumably located in the areas with the highest activity. However, thanks to the local weighting of the bandwidth factors, the adaptive KDE is able to better isolate highly active areas, with less probability mass concentrated in the entrance of the port.

Finally, Fig. 4 demonstrates the effect of the parameter $\alpha$ on the resulting estimate. It is apparent how smaller values of $\alpha$ tend to produce similar results as the fixed KDE, while higher values make the density estimate spikier but necessarily more fragmented.

## V. Conclusion and future work

Estimating port locations and operational areas is an essential component for achieving MSA. The large volume of AIS data imposes algorithmic approaches that require minimal human intervention and scale with the increasing data volumes. The KDE-based approaches presented here address these challenges by combining MapReduce with fixed or adaptive kernel bandwidths. The results presented on the single port of Rotterdam could be extended to other ports worldwide, and a port analysis platform could be developed that learns the port areas worldwide in an unsupervised way. The proposed approach can also be extended to other types of areas besides ports: off-shore platforms, anchorage areas, and fishing grounds can be detected automatically and their extent estimated in a data-driven, unsupervised fashion.

## References


[1] EU Commission, "Ports 2030 – Gateways for the trans European transport network," 2014.

[2] ——, "Ports: An engine for growth," 2013, COM(2013) 295 final.

[3] H. Haralambides and M. Acciaro, "The new European port policy proposals: too much ado about nothing?" *Maritime Economics & Logistics*, vol. 17, no. 2, pp. 127–141, 2015.

[4] EU Commission, "Regulation of the European Parliament and of the Council establishing a framework on market access to port services and financial transparency of ports," 2016.

[5] European Sea Ports Organisation (ESPO), "Annual report 2014–2015," 2015.

[6] H. Haralambides, "Competition, excess capacity, and the pricing of port infrastructure," *International journal of maritime economics*, vol. 4, no. 4, pp. 323–347, Dec. 2002.

[7] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big Data*, vol. 1, no. 1, pp. 51–59, Feb. 2013.

[8] C. Ducruet, C. Rozenblat, and F. Zaidi, "Ports in multi-level maritime networks: evidence from the Atlantic (1996–2006)," *Journal of Transport Geography*, vol. 18, no. 4, pp. 508–518, 2010.

[9] M. Tichavska, F. Cabrera, B. Tovar, and V. Araña, "Use of the Automatic Identification System in academic research," in *International Conference on Computer Aided Systems Theory*. Springer, 2015, pp. 33–40.

[10] G. Pallotta, M. Vespe, and K. Bryan, "Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction," *Entropy*, vol. 15, no. 6, pp. 2218–2245, Jun. 2013.

[11] B. Huijbrechts, M. Velikova, S. Michels, and R. Scheepens, "Metis: An integrated reference architecture for addressing uncertainty in decision-support systems," *Procedia Computer Science*, vol. 44, pp. 476–485, 2015.

[12] M. Rashidi and J. Koto, "Prediction of CO2 emitted by marine transport in batam-singapore channel using AIS," *Jurnal Teknologi*, vol. 69, no. 7, Jul. 2014.

[13] B. Ristic, B. La Scala, M. Morelande, and N. Gordon, "Statistical analysis of motion patterns in AIS data: Anomaly detection and motion prediction," in *2008 11th International Conference on Information Fusion*, Jun. 2008, pp. 1–7.

[14] S. Ricci, C. Marinacci, and L. Rizzetto, "The modelling support to maritime terminals sea operation: The case study of port of Messina," *Journal of Maritime Research*, vol. 9, no. 3, pp. 39–44, 2014.

[15] A. Donateo, E. Gregoris, A. Gambaro, E. Merico, R. Giua, A. Nocioni, and D. Contini, "Contribution of harbour activities and ship traffic to PM2.5, particle number concentrations and PAHs in a port city of the Mediterranean Sea (Italy)," *Environmental Science and Pollution Research*, vol. 21, no. 15, pp. 9415–9429, Apr. 2014.

[16] L. V. Jing, "Safety evaluation of China's maritime transport key nodes," *Journal of Transportation Systems Engineering*, vol. 15, no. 1, pp. 30–36, Dec. 2014.

[17] International Maritime Organization, "International convention for the Safety Of Life At Sea (SOLAS)," 2003.

[18] B. Tetreault, "Use of the Automatic Identification System (AIS) for maritime domain awareness (MDA)," in *OCEANS, 2005. Proceedings of MTS/IEEE*, Sept 2005, pp. 1590–1594 Vol. 2.

[19] P. Last, C. Bahlke, M. Hering-Bertram, and L. Linsen, "Comprehensive analysis of Automatic Identification System (AIS) data in regard to vessel movement prediction," *The Journal of Navigation*, vol. 67, pp. 791–809, 9 2014.

[20] G. Cimino, G. Arcieri, S. Horn, and K. Bryan, "Sensor data management to achieve information superiority in Maritime Situational Awareness," NATO STO CMRE, Tech. Rep. CMRE-FR-2014-017, 2014.

[21] M. Rosenblatt *et al.*, "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, 1956.

[22] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[23] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26.

[24] ——, "Algorithm AS 176: Kernel density estimation using the fast Fourier transform," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 31, no. 1, pp. 93–99, 1982.

[25] Y. Chen, M. Welling, and A. Smola, "Super-samples from kernel herding," *arXiv preprint arXiv:1203.3472*, 2012.

[26] J. M. Phillips, B. Wang, and Y. Zheng, "Geometric inference on kernel density estimates," *arXiv preprint arXiv:1307.7760*, 2013.

[27] Y. Zheng, J. Jestes, J. M. Phillips, and F. Li, "Quality and efficiency for kernel density estimates in large data," in *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, 2013, pp. 433–444.

[28] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[29] National Geospatial-Intelligence Agency (NGA), "World Port Index (Pub 150)," Springfield, Virginia, 2016, twenty-fifth ed.