A distributed approach to estimating sea port operational regions from lots of 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 play a vital role in the global economy, as they operate as the connection corridors to all other modes of transport and as engines of growth for the wider region. But ports today are faced with numerous unique challenges and for them to remain competitive, significant investments are required. In support of greater transparency in policy making, decisions regarding investment need to be supported by data-driven intelligence. It is often an overlooked fact that seaports do not remain static over time; such spatial units often evolve according to environmental patterns both in size but also connectivity 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. In this work, we leverage the huge amounts of vessel data that are progressively becoming available through the Automatic Identification System (AIS) and distributed machine learning to define a seaport's extended area of operation. Specifically, we present our adaptation of the well-known KDE algorithm to the map-reduce paradigm, and report results on the port of Shanghai.**

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

I. Introduction

Today, more than 80% of world trade is transported by sea. Over the years, the shipping industry has often had to adapt to market volatility and economic instability. According to Alphaliner Research, in 2015, a record 212 new container ships were delivered, increasing the global fleets total capacity by 8.5 percent. Simultaneously there was a noticeable trend towards higher capacity ships; a trend which has accelerated in the past five years. Vessels have grown drastically in size so as to improve fleet overall efficiency, allowing fewer sailings to the same amount of transportation units, while they are becoming safer and greener so as to adapt to stricter safety and environmental requirements.

On the flip side, terminal operators and port authorities have been unable to adapt to the rapidly changing conditions in their ports; as such terminals have struggled to handle the growing volumes of containers especially during shipping peaks and traffic jams have generated long delays across all modes of transport. For example, in 2013 Los Angeles and Long Beach, which together handled 41% of US container traffic in 2013, were full to overflowing during the peak months of August to December, with sometimes up to 18 vessels having to wait outside the ports. New cranes, taller bridges, terminals and even reconfiguration of the container yards are just some of the costly investments required by ports to receive these huge vessels and service them efficiently. Although investments are usually funded by the government or other public bodies, returns on investments can be made by higher port fees but huge ships are making fewer port calls, while each call is more expensive.

Similarly to other industries, decisions regarding the re-design of ports areas, their increase in operational capacity and infrastructure, need to be based on measurable data which can be transformed into actionable information.

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. 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. 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 specific categories of ships are required to carry, that is a collaborative, self-reporting system that allows vessels to broadcast their identity, position, and other information to nearby vessels and on-ground base stations.

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

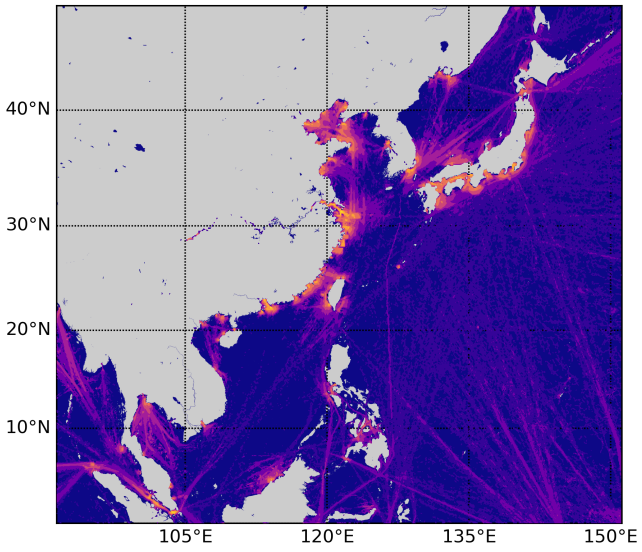


Figure 1. Density of AIS messages collected by MarineTraffic during March 2015. Each pixel covers a 6-by-6 nmi (one-tenth-degree) square on the ground and its color is (logarithmically) proportional to the number of AIS messages whose reported positions fall within its footprint.

such 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 [1]. The stepping stone for any useful analytics and valid data driven approach to port planning is accurately defining a seaport's location and operational boundaries, so capacity and efficiency can be calculated.

II. Background and related work

AIS data analysis has proven to be a valid method for monitoring vessels and extracting valuable information regarding vessel behaviour, 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 [2]. Pallotta, Vespe and Bryan make use of AIS data for vessel pattern knowledge discovery as a framework for anomaly detection and route prediction [3]. In [4], 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 seaports research and AIS, in [5] AIS is leveraged to model maritime terminals operations, specifically focusing on the port of Messina, Italy, while in [6] is evaluated the contribution of harbor activities and ship traffic to PM2.5, particle number concentrations and PAHs in anoter port city of the Mediterranean Sea (Italy) [6].

Interestingly though, only a small number of these publications describe applications of MapReduce and Hadoop approaches to the maritime domain, although there is an apparent need for parallel computation due to the enormous amount of implicated data. In their work [7], Wang et al. attempt to tackle the big data issue caused by the AIS data for anomaly detection purposes. They implement a two-step process, where they firstly use an unsupervised technique, based upon the Density-Based Spatial Clustering of Applications with Noise considering Speed and Direction (DBSCAN-SD) incorporating non-spatial attributes, such as speed and direction, to label normal and abnormal position points of vessels based on the raw AIS data. Secondly, they train a supervised learning algorithm designed with the MapReduce paradigm running on Hadoop using the labeled data generated in from the first step. HBase, a distributed scalable column-oriented database part of the Apache Hadoop ecosystem, is used in [8] to store, process, and analyze a large amount of spatiotemporal data generated by shipboard AIS transponders and following this a simple method of predicting vessel behavior is proposed.

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. In this paper we present a methodology to estimate port locations and operational areas in a scalable and unsupervised way, based on Kernel Density Estimator (KDE).

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

In the remainder of this paper, we apply our approach to a dataset of more than 57 million AIS messages made available by MarineTraffic and recorded during the month of March 2015, in an area of interest that spans more than 32×10^6 square km, approximatively from $90°$ to $150°$ longitude and from $0°$ to $50°$ latitude. 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 6-by-6 nmi (one-tenth-degree) 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.

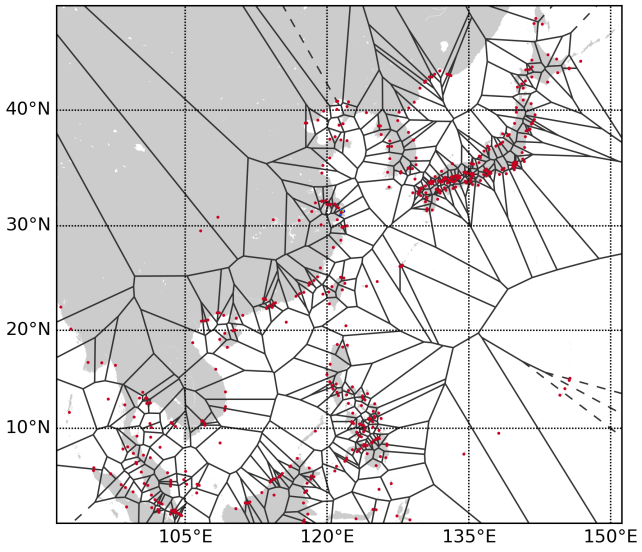


Figure 2. The WPI enumerates 483 ports in the area of interest that belong to 23 different countries. The port locations are shown on the map with red dots. A Voronoi diagram is also overlaid that depicts how the 1-Nearest Neighbor (1-NN) splits the data in different partitions.

B. Data preparation

The first hurdle is related to the fact that, in order for the KDE to be meaningful, only the positions of ships moored in the port of interest should be taken into account. Unfortunately, this information is not available *a priori*, but has to be determined from the data. The World Port Index (WPI) [9] is an open database, maintained and updated by the US National Geospatial-Intelligence Agency (NGA) that contains the location and physical characteristics of, and the facilities and services offered by major ports and terminals world-wide (approximately 3700 entries), in a tabular format. Entries are organized geographically, in accordance with the diagrams located in the front of the publication.

The WPI enumerates 483 ports in the area of interest that belong to 23 different countries, as also depicted in Fig. 2, where red dots indicate the locations of the ports. Using this information, we use a simple k-Nearest Neighbor (k-NN) algorithm [10] to determine the closest port —$k = 1$, or 1-NN— for each position reported by the AIS. The Voronoi diagram depicted in Fig. 2 is a graphical representation of this concept, being each region defined as the set of points in the space that are closer to the seed (i.e. the position of the corresponding port) of that region than to any other seed. In other words, the 1-NN enables us to create logical partitions of the input data set that can be used afterwards to construct the KDE for each port in the area of interest.

In this work we focus on the port of Shanghai, as this is the busiest container port in the world and with one of the most complex operational areas to identify. The port of Shanghai includes 3 major working zones: the Yangshan Deep-Water Port (not visible in the figures), the Huangpu River and the Yangtze River. This port is the main transport hub for foreign trade in the area. Within the context of this work, we exploit the fact that vessels slow down when entering port areas, before coming to a complete stop. We focus on vessels with speed below $1\,\mathrm{kn}$ and on their positional information (longitude and latitude).

C. 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 [11], a basic KDE of f has the form [12]:

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

where K_h is the kernel function, and h is the 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 [13]. 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 a similar result at a fraction of the computational cost.

2) Adaptive KDE: Both the KDE in (1) and the KDE by convolution 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 [13] 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 of how the density is distributed in the area of interest, here we can use the convolved histogram, which comes at a fraction of the computational cost required to evaluate (1).

3) Local bandwidth factors: Under the assumption that the underlying distribution is k-variate normal, the optimum (fixed) window can be written [13] as $h^* = \left(\frac{4}{n(k+2)}\right)^{\frac{1}{k+4}}$. The *local bandwidth* factors λ_i, for $i = 1, \ldots, n$ are then given by $\lambda_i = \left(\frac{f_n(\boldsymbol{x}_i)}{g}\right)^{-\alpha}$, where $0 \leq \alpha \leq 1$ is the sensitivity parameter and g is the geometric mean of the fixed-bandwidth density estimate $f_n(\boldsymbol{x}_i)$ evaluated in the data points, i.e. $\log g = \frac{1}{n}\sum_{i=1}^{n} \log f_n(\boldsymbol{x}_i)$. The adaptive KDE of f can be finally expressed as

$$\hat{f}_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(h^*\lambda_i)^k} K_{h^*\lambda_i}(\boldsymbol{x}, \boldsymbol{x}_i). \tag{3}$$

IV. Implementation and results

Let us indicate the kinematic state of a vessel at a generic time with $\boldsymbol{x}_i = [a_i, b_i]^{\mathsf{T}} \in \mathbb{R}^2$, where a and b represent the longitude and latitude coordinates, respectively, of the ship in a geographic coordinate system. 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 positional AIS observations $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ 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 [14]–[16]. 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. [17] have recently proposed randomized and deterministic distributed algorithms for efficient KDE with quality guarantees, adapting them to the popular MapReduce programming model. As in [17], 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. 3 we report a conceptual representation of the formulation of the kernel density estimation problem in the MapReduce framework. The leftmost blocks represent the partitions of the input data relative to a single port. As already mentioned in Sect. III-B, this can be achieved using a 1-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. We form the KDE in $\mathbb{R}^2$ using 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.

Filtering out all the ships whose velocity exceeds the threshold of $1\,\mathrm{kn}$ leaves us with a dataset of ≈ 19 million samples, from an initial size of ≈ 57 million. These ≈ 19 million data samples are then fed into the 1-NN classifier to identify the nearest port to each samples. The result of this operation is stored in an intermediate data frame that enables us to select the data samples that should contribute to the construction of the KDE for the given port of interest. In the case under study, for the port of Shanghai, ≈ 1.8 million positions have been found to be considered for computing the density estimate, corresponding to the AIS messages received by MarineTraffic during 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 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, finding the nearest port to each data sample takes about $7\,\mathrm{minutes}$, while the fixed KDE on the port of Shanghai takes about $6\,\mathrm{minutes}$. 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 $12\,\mathrm{minutes}$ to run.

In Fig. 4 we report the comparison between the fixed-bandwidth (a) and adaptive (b) KDE computed in the area of the port of Shanghai, the most trafficked port in the area of interest. The estimate has been determined using the available data collected by MarineTraffic during March 2015, 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 WPI, and most of the probability mass is in both cases concentrated the area of the Yangtze River. Another significant part the PDF produced by the fixed-bandwidth KDE follows continuously the Huangpu River, and exhibits four distinct peaks around the areas with the greatest activity. Thanks to the local weighting of the bandwidth factors, the adaptive KDE is able to better isolate highly active areas along the Huangpu River, with less probability mass concentrated in the Yangtze River.

Finally, Fig. 5 demonstrates the effect of the parameter α on the resulting estimate. It is apparent how smaller values of α tend to produce similar results as the fixed KDE,

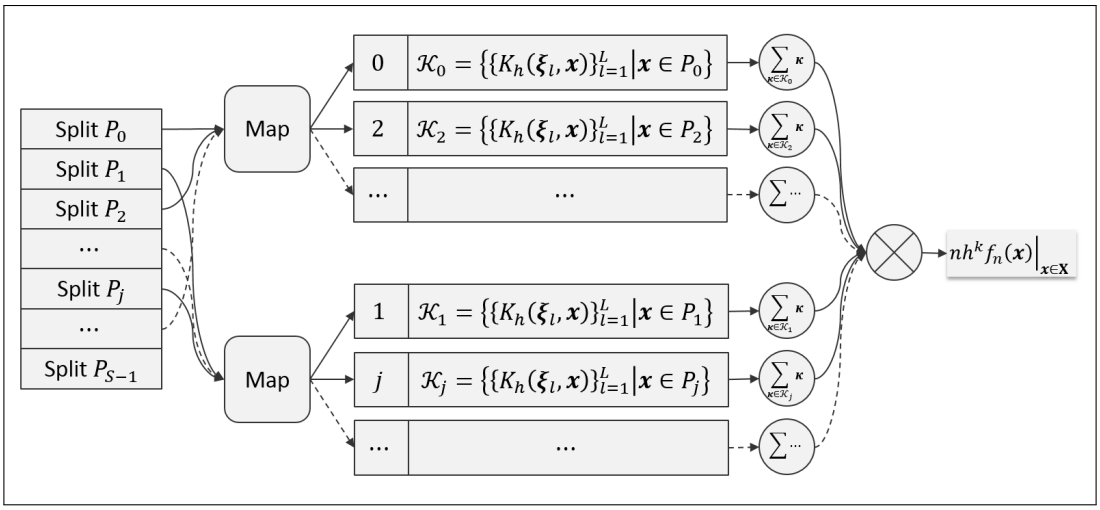
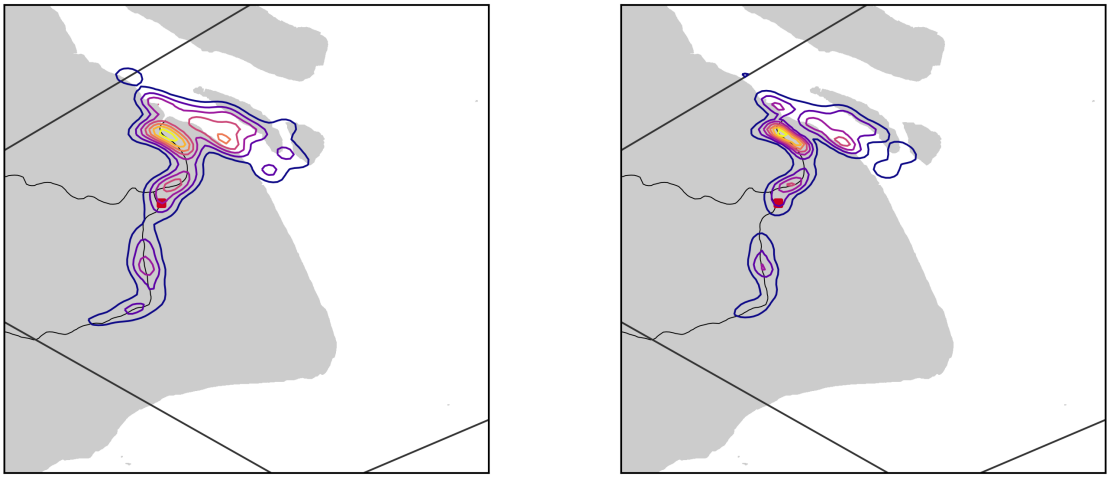


Figure 3. Implementation of the KDE with MapReduce logic. The leftmost block shows the the input data relative to a single port, which is organized in partitions and stored on a DFS or a distributed database. Taking advantage of its linearity, the computation of the KDE can be distributed among multiple nodes, each of which performs an expansion of the input partition with the Gaussian kernel.



(a) Fixed-bandwidth KDE

(b) Adaptive KDE

Figure 4. Comparison of fixed and adaptive bandwidth kernel density estimates computed in the Shanghai 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 during March 2015, having selected only those ships whose speed reported by the AIS was not exceeding the fixed threshold of 1 kn. The red square marker in the map shows the position of the port as recorded in the WPI.

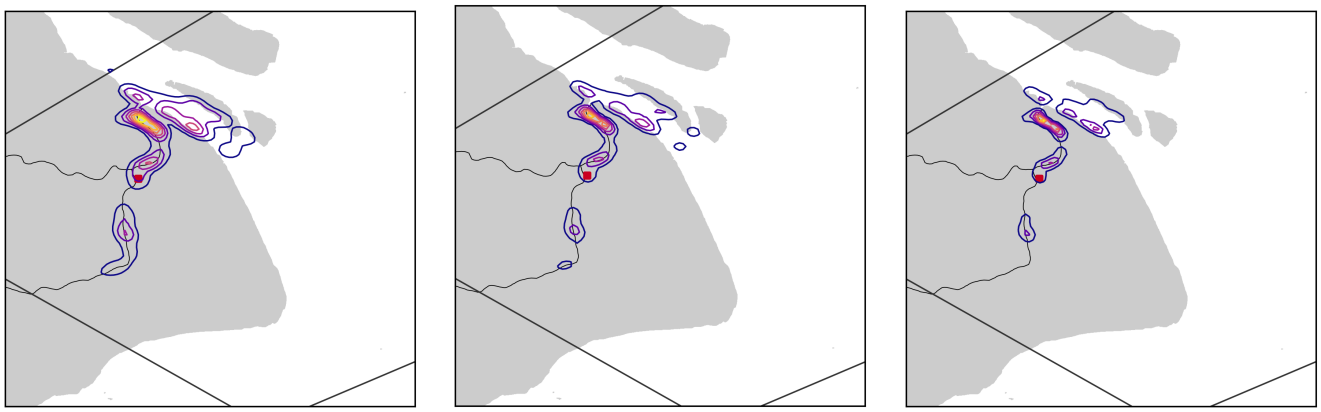
while higher values make the density estimate spikier but necessarily more fragmented.

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 Shanghai 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 be extended to other types of areas besides ports, to automatically estimate their extent in a data-driven, unsupervised fashion.

Acknowledgment

This work is supported by the NATO SACT under project SAC000608, by Microsoft Research through a Microsoft Azure for Research Award, and by MarineTraffic.



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

Figure 5. Effect of the sensitivity α 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 α produce results that are more similar to the fixed-bandwidth KDE, while *spikier* density functions are created with higher values of α, but they are inevitable more fragmented.

REFERENCES

[1] C. Ducruet and C. R. amd F. Zaidi, "Ports in multi-level maritime networks: evidence from the Atlantic (19962006)," *Journal of Transport Geography*, vol. 18, no. 4, pp. 508–518, Jul. 2010.

[2] M. Tichavska, F. Cabrera, B. Tovar, and V. Araa, "Use of the Automatic Identification System in Academic Research," in *EUROCAST 2015*, ser. Lecture Notes in Computer Science, R. Moreno-Daz, F. Pichler, and A. Quesada-Arencibia, Eds. Springer, Feb. 2015, no. 9520, pp. 33–40.

[3] 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. [Online]. Available: http://www.mdpi.com/1099-4300/15/6/2218

[4] B. Ristic, B. L. 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.

[5] S. Ricci, C. Marinacci, and L. Rizzetto, "The Modelling Support to Maritime Terminals Sea Operation: The Case Study of Port of Messina," 2014.

[6] 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.

[7] X. Wang, X. Liu, B. Liu, E. de Souza, and S. Matwin, "Vessel route anomaly detection with Hadoop MapReduce," in *Big Data (Big Data), 2014 IEEE International Conference on*, Oct 2014, pp. 25–30.

[8] W. Wijaya and Y. Nakamura, "Predicting ship behavior navigating through heavily trafficked fairways by analyzing AIS data on Apache HBase," in *Computing and Networking (CANDAR), 2013 First International Symposium on*, Dec 2013, pp. 220–226.

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

[10] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[11] 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.

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

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

[14] ——, "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.

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

[16] J. M. Phillips, B. Wang, and Y. Zheng, "Geometric inference on kernel density estimates," *CoRR*, vol. abs/1307.7760, 2013.

[17] 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.