SCALABLE ESTIMATION OF PORT AREAS FROM AIS DATA

Leonardo M. Millefiori, Luca Cazzanti, Dimitris Zissis, Gianfranco Arcieri

NATO STO-CMRE Centre for Maritime Research and Experimentation
Viale San Bartolomeo, 400
La Spezia, Italy

ABSTRACT

This paper discusses work in progress to estimate port locations and operational areas in a scalable, data-driven, unsupervised way. Knowing the extent of port areas is an important component of larger maritime traffic analysis systems that inform stakeholders and decision makers in the maritime industry, governmental agencies, and international organizations. The proposed approach uses Kernel Density Estimator (KDE) and exploits the large volume of Automatic Identification System (AIS) data to learn the extent of port areas in a data-driven way. Example results for the port of La Spezia, Italy, demonstrate the approach for real data.

Index Terms— KDE, port location estimation, AIS, map reduce, big data

1. INTRODUCTION

The Mediterranean hosts one of the most complex and dense port networks in the world, a gateway to European commerce and industry. A recent report published by the European Commission calculated that 74% of goods imported and exported and 37% of exchanges within the European Union transit through the roughly 1,200 seaports along its 70,000 km of coastline [1, 2]. Decision makers, policy advisors, trade partners, security experts, safety agencies, international organizations, and vessel operators are becoming more reliant on benchmark metrics of port activities to carry out their duties. Examples of such metrics include maritime and intermodal connectivity indicators, volume of cargo throughput, proportion of different types of goods transported, and fishing activity indicators. In addition, stakeholders are becoming more reliant on summary statistics and representative Patterns of Life (PoLs) to characterize the ports according to their local and regional traffic patterns and operational capabilities.

Generating valid and reliable measurements though, is a complex task. We often overlook the fact that maritime networks operate as "small worlds", where content and size vary over space and time, under the influence of the trade and carrier patterns. In particular, port region, port system, and port range are spatial entities that evolve over time [3], yet their clear definitions are essential for obtaining accurate metrics on port activities.

Manually collating and curating port area definitions is not a realistic approach: subjective definitions of port areas and system maintenance costs make it unreliable and infeasible. Thus, the stepping stone for any useful port analysis is an automatic, unsupervised, data-driven approach to defining seaport locations and operational boundaries. While in the past, sea transport surveillance had suffered from a lack of data, current tracking technology such as Automatic Identification System (AIS) [4] has transformed the problem into one of extracting interpretable information from an overabundance of maritime data. This introduces the additional requirement that the algorithmic approaches must be scalable to big data regimes.

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, have become available. These datasets provide detailed insights into the patterns vessels follow, while they can operate as benchmarking tools for port authorities regarding the effectiveness and efficiency of their ports.

To address the big data challenge in the maritime domain, researchers have developed computational and statistical approaches that rely on AIS data to automatically monitor vessel activities and extract their behavioral patterns [5, 6, 7, 8]. Previous work at the NATO Science and Technology Organization Centre for Maritime Research and Experimentation (STO-CMRE) has explored how to extract stationary areas from AIS data based on the spatial clustering algorithm DBSCAN [9, 10]. Building on that initial work, this paper presents work-in-progress to estimate port areas in a scalable, unsupervised, data-driven way. Instead of DBSCAN, the approach relies on the Kernel Density Estimator (KDE) to form

This work is supported by the NATO Supreme Allied Command Transformation (SACT) under the project SAC000608 – Data Knowledge Operational Effectiveness.

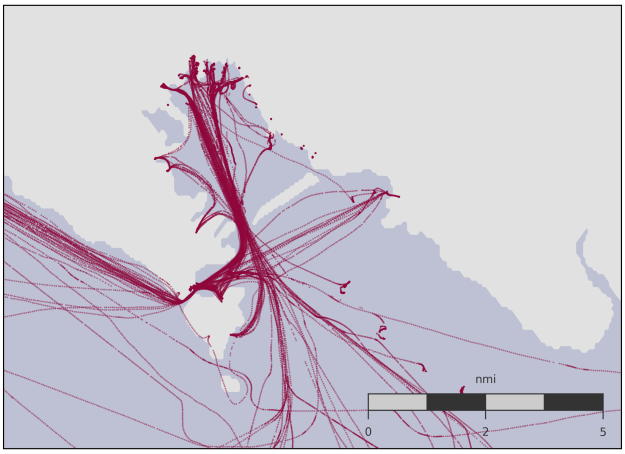

Fig. 1. Maritime traffic in the port of La Spezia in a 24-hour period during August 2015. The ship positions are received by CMRE's local AIS receiver.

density-based estimates of port locations and operational areas from the location and velocity data contained in the AIS messages transmitted by the vessels. The approach is scalable because the operations underlying KDE are decomposable into MapReduce primitives [11, 12], which enables distributing the computational load across different computing nodes and across distributed storage.

2. 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 [13], a basic KDE of f has the form [14]:

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

where K_h is the kernel function, and h denotes the 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

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

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 [15], 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 ϕ_n is the empirical PDF, expressed as a sum of n Dirac delta functions $\delta\left(\cdot\right)$ 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.

2.2. 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 [15] 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 (3), which comes at a fraction of the computational cost required to compute (1).

2.2.1. *Local bandwidth factors*

Under the assumption that the underlying distribution is k-variate normal, the optimum (fixed) window can be written as [15]:

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

The *local bandwidth* factors λ_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

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

3. IMPLEMENTATION AND RESULTS

Let us indicate 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 AIS of observations $\mathcal{X}$, 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 [16, 17, 18]. 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. [12] have recently proposed randomized and deterministic distributed algorithms for efficient KDE with quality guarantees, adapting them to the popular MapReduce programming model. As in [12], our approach is to take advantage of the linearity of the KDE to distribute the computation among many different nodes using the *MapReduce* [11] distributed programming model.

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. The first one is to compute the KDE in $\mathbb{R}^3$ at a 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.

The second approach 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 second approach can be also seen as an approximation to the first one that trades some result accuracy for a more affordable computational cost.

We applied these two approaches and the adaptive KDE (7) to the data set shown in Fig. 1, which is made up by all the AIS messages received by CMRE's local station during a 24-hour period in August 2015. The resulting kernel density estimates are shown in Fig. 2, where we report: on the left side, the fixed-bandwidth KDE computed in $\mathbb{R}^3$ using the kinematic states $\boldsymbol{\chi}_i$ and marginalized on v; in the middle and on the right side, the fixed-bandwidth and adaptive KDE, respectively, both computed in $\mathbb{R}^2$ and discarding all kinematic states whose instantaneous speed was greater than the threshold v_T, that was set, in all three cases, to $1\,\mathrm{kn}$.

4. CONCLUSION AND FUTURE WORK

Estimating port locations and operational areas is an essential component for achieving Maritime Situational Awareness (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 La Spezia 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: off-shore platforms, anchorage areas, and fishing grounds can be detected automatically and their extent estimated in a data-driven, unsupervised fashion.

5. REFERENCES

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

[2] ——, "Ports: an engine for growth, COM(2013) 295 Final," Tech. Rep., 2013.

[3] C. Ducruet and C. R. amd 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, Jul. 2010.

[4] International Telecommunications Union, "Technical characteristics for an automatic identification system using time division multiple access in the VHF maritime mobile band (Recommendation ITU-R M.1371-4)," 2012.

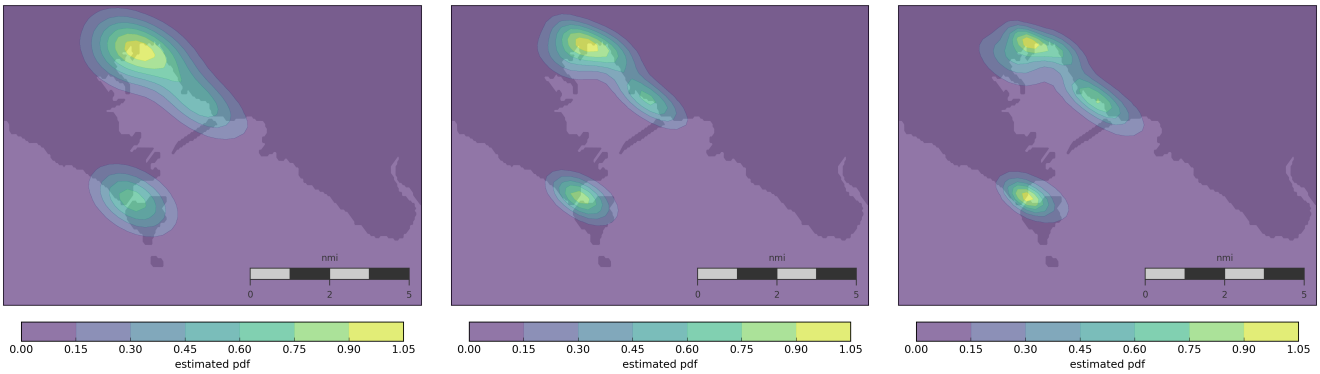

Fig. 2. Comparison of kernel density estimates computed with different approaches: fixed bandwidth with $\boldsymbol{x}_i \in \mathbb{R}^3$ (left) and $\boldsymbol{x}_i \in \mathbb{R}^2$ (middle), and adaptive bandwidth with $\boldsymbol{x}_i \in \mathbb{R}^2$. The fixed $\mathbb{R}^3$ version produces the smoothest result, but is unable to deal satisfactory with the low-density estimate regions, and has the highest computational cost. The fixed approach in $\mathbb{R}^2$ is computationally more affordable, but is equally not able to produce satisfactory results in low-density regions. Finally, the adaptive KDE in $\mathbb{R}^2$ on the right has a higher computational cost than the fixed KDE, but it is the only one that produces a *spikier* estimate on low-density regions. All the three estimates have been computed on the AIS messages received by CMRE's local base station in a 24-hour time span during August 2015, shown in Fig. 1. The speed threshold that discriminates stationary from non-stationary targets is set to $1\,\mathrm{kn}$.

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

[6] 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

[7] 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.

[8] F. Mazzarella, M. Vespe, D. Damalas, and G. Osio, "Discovering vessel activities at sea using AIS data: Mapping of fishing footprints," in *Information Fusion (FUSION), 2014 17th International Conference on*, July 2014, pp. 1–7.

[9] L. Cazzanti and G. Pallotta, "Mining maritime vessel traffic: Promises, challenges, techniques," in *OCEANS*. Genova, Italy: IEEE/MTS, May 2015.

[10] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Second International Conference on Knowledge Discovery and Data Mining*, Portland, OR-US, 1996.

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

[12] 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.

[13] 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.

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

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

[16] ——, "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.

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

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