

# Shape-Based Analysis for Vessel Trajectories

Jiang Wang, Yun Zhou, Xiaofeng Cao, Yilin Wang, Cheng Zhu, and Weiming Zhang  
Science and Technology on Information Systems Engineering Laboratory,  
National University of Defense Technology, 410073, Changsha, China  
wang1988jiang@foxmail.com, mekiddcxf@126.com,  
{zhouyun, wangyilin15, zhucheng, zhangweiming}@nudt.edu.cn

## ABSTRACT

In this paper we propose a novel method for modeling the shape of vessel trajectories in a manner which may facilitate the application of machine learning techniques. This is achieved by transforming the topological feature of vessel trajectories into vectors. More specifically, we calculate scale-invariance indicators for every vessel trajectory as shape characteristics, and other indicators to denote the trajectory area. The proposed method is validated using both synthetic trajectories and real-world AIS datasets. We demonstrate that it can achieve good time efficiency and may support vessel trajectory related analysis.

## CCS CONCEPTS

• **Computing methodologies** → **Spatial and physical reasoning**;  
• **Information systems** → **Spatial databases and GIS**.

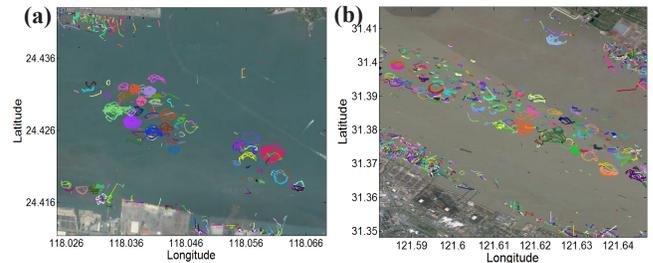
## KEYWORDS

Trajectory shape, Automatic identification system, Spatial data, Scale-invariance

## 1. INTRODUCTION

Large volumes of data produced by the Automatic Identification System (AIS) [1] provide opportunities for studying vessel mobility patterns. Some works study the topology or self-similar patterns in land traffic data [2, 3]. However, unlike land moving objects such as vehicles, human or animals, vessels show different move and stop patterns, e.g., move across a wide range on open sea and sail along rivers on inner waters. A vessel may follow the direction of water flow and wind when engine is turned off. Fig. 1 shows trajectories of vessels at anchor or moored, in Xiamen Bay and the Yangtze River, respectively. Some moored vessels follow circle-like trajectories, which are obviously different from vehicles that report overlapping location points when they stop.

As with the peripheries of clouds and the coastlines [4], we observe the majority of real vessel trajectories show a property of scale-invariance, namely they are statistical self-similar and exhibit fractals like behaviors. This provides a way for modeling



**Figure 1: Trajectories of vessels at anchor or moored. (a) Xiamen Bay, China. (b) The Yangtze River (Shanghai, China).**

the shape of vessel trajectories, without considering their geographical locations. Shape-based trajectory analysis could be applied to many maritime related tasks, such as: (i) find ship mobility patterns on temporal and spatial dimensions; (ii) analyze trajectory shapes derived from ships with diverse types and displacements; (iii) transform massive amounts of raw AIS data into high-level knowledge by machine learning techniques, e.g., shape-based trajectory clustering and classification.

In order to automatically and efficiently identify vessel behaviors, we quantitatively examine the self-similarity characteristic of vessel trajectories from the view of fractal geometry. A novel approach is proposed for calculating scale-invariance indicators to describe the trajectory shape. We show how the indicators support vessel trajectories related analysis. Our contributions are in three folds: (i) To our awareness, this is the first work to examine and provide theoretical analysis on the self-similarity characteristic of vessel trajectories. (ii) An efficient method is developed to extract the scale-invariance feature of vessel trajectory shape, which is linear to the number of points of trajectory. (iii) We perform experiments on synthetic data to validate and on real-world datasets to identify vessel behaviors.

## 2. TRAJECTORY ROTATION AND PROJECTION

*Definition 1. (Major Axis)* Given the geometric center  $O$  of a trajectory, and let  $d_{ij}$  be the Euclidean distance between points  $P_i$  and  $P_j$ , the major axis  $MA$  of the trajectory is defined as  $OP_A$ , where  $A$  is calculated by  $A = \arg \max (d_{OP_i})$ .

Once the major axis is determined, other axes can be defined based on it. For instance, use its perpendicular line as the minor axis  $MI$ .

*Definition 2. (Trajectory Rotation)* A trajectory can be rotated around  $MA$  or  $MI$  or other defined axes, and we call this process as trajectory rotation.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

SIGSPATIAL'17, November 7–10, 2017, Los Angeles Area, CA, USA

© 2017 Copyright is held by the owner/author(s).

ACM ISBN 978-1-4503-5490-5/17/11.

<https://doi.org/10.1145/3139958.3140060>

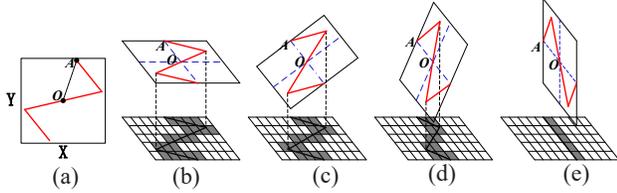


Figure 2: The process of trajectory rotation and projection.

*Definition 3. (Projective Trajectory)* In the process of trajectory rotation, for each rotation angle  $\theta_j, j = 1, 2, \dots, J$ , there's a projective trajectory on the plane that parallels to the rotation axis.

Fig. 2 illustrates the process of trajectory rotation and projection using a synthetic trajectory. In the original coordinate system, geometry center  $O$  and point  $A$  which has the maximal distance from  $O$  are computed, as shown in Fig. 2(a). Then new coordinates are computed for each point in the new coordinate system, and the trajectory is projected onto a horizontal plane, which is partitioned by  $m$  uniform grids, as shown in Fig. 2(b). Fig. 2(b)-2(e) show that the trajectory is rotated around  $OA$  along the direction of counterclockwise. Meanwhile, the non-empty grids  $Ng$  are counted for each projective trajectory. Points on the borderlines are assigned to the right grids.

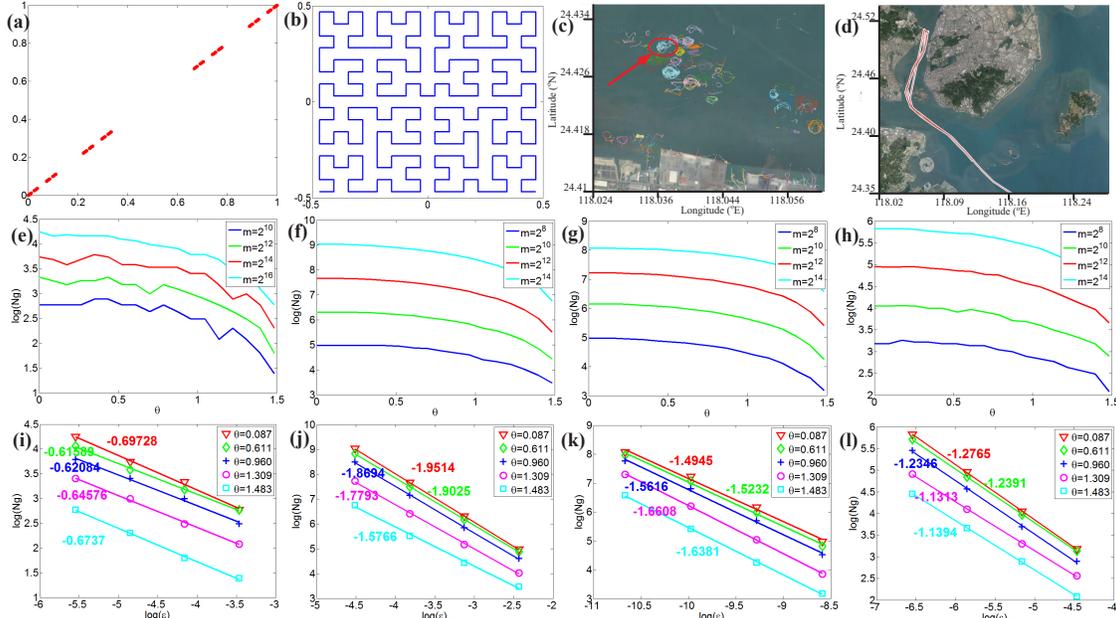


Figure 3: (a)-(d) Cantor dust, Hilbert curve, a vessel stop trajectory (in circle) and a move trajectory. (e)-(h)  $\log(Ng)$  vs.  $\theta$  under different scales for the 4 datasets, the rotation axis is  $MI$ , along the direction of counterclockwise. (i)-(l)  $\log(Ng)$  vs.  $\log(\epsilon)$  under different  $\theta$ s for the 4 datasets, the slope of each line is  $-FD$ , and  $\epsilon$  is the grid length under scales.

### 3. PROPERTY OF SCALE-INVARIANCE

Fig. 3(a)-3(d) show 4 datasets, the first two are the Cantor dust and the Hilbert curve; datasets 3 and 4 are real vessel trajectories, in which one is a stop and the other is a move. The Cantor dust is derived by recursively deleting the middle third of a line segment and the Hilbert curve is a space filling curve. Both the two are fractals and well known for their self-similar property [5], and have fractal dimensions  $FD_{Cantor} \approx 0.63$  and  $FD_{Hilbert} = 2$ .

Using rotation and projection method (around  $MI$ ) to count  $Ng$  for each dataset under different scales (grid number  $m$ ), the resulting curves are depicted in Fig. 3(e)-3(h). As these curves have periodicity with respect to the rotation angle  $\theta$ , they are shown in  $\theta \in (0, \pi/2)$ . The symbol 'log' means the natural logarithm in this paper (i.e., with base  $e$ ).

**OBSERVATION 1.** Like the Cantor dust and the Hilbert curve, the  $\log(Ng(\theta))$ - $\theta$  curves of a real vessel trajectory under different scales have the same trend.

Namely, if fit  $\log(Ng(\theta))$ - $\theta$  curves of a trajectory with  $S$  quadratic functions  $\log(Ng_{s_i}(\theta)) = a_{s_i}\theta^2 + b_{s_i}\theta + c_{s_i}, i = \{1, 2, \dots, S\}$ , then

$a_{s_1} \approx a_{s_2} \approx \dots \approx a_{s_S}$ , and  $b_{s_1} \approx b_{s_2} \approx \dots \approx b_{s_S}$ , they are only different in the constants  $c_{s_i}$ . This observation is true for the vast majority of vessel trajectories, as well as using  $MA$  as the rotation axis.

**OBSERVATION 2.** For a real vessel trajectory, its projective trajectories share a similar fractal dimension  $FD$ .

In Fig. 3(i)-3(l), the fitted  $\log(Ng)$ - $\log(\epsilon)$  lines of the projective trajectories (namely, different  $\theta$ s) of each dataset share a similar slope, and this observation is also true for other vessel trajectories. Then we assume that the projective trajectories of a self-similar trajectory are self-similar and share a same  $FD$ .

**LEMMA 1.** If a trajectory is statistical self-similar, then the fitted coefficients  $\mathbf{a}$  and  $\mathbf{b}$  have the property of scale-invariance, and vice versa.

**PROOF.** For a statistical self-similar point-set in  $n$ -dimensional space, the (Hausdorff) fractal dimension  $FD$  satisfies the exponent of the law [6]:  $Ng(\epsilon) = C * \epsilon^{-FD}$ , ( $r_1 < \epsilon < r_2$ ), where  $C$  is a constant and  $\epsilon$  is the grid length within a range of scales ( $r_1, r_2$ ). Then for scales  $s_i$  and  $s_j$ , ( $i, j = 1, 2, \dots, S$ ),  $\log(Ng(\epsilon_{s_i})/Ng(\epsilon_{s_j})) = -FD * \log(\epsilon_{s_i}/\epsilon_{s_j})$ , which is a constant.

Let  $\log(Ng_{si}(\theta)) = a_{si}\theta^2 + b_{si}\theta + c_{si}$  and  $\log(Ng_{sj}(\theta)) = a_{sj}\theta^2 + b_{sj}\theta + c_{sj}$  be two fitted quadratic functions under scales  $si$  and  $sj$  respectively. With the assumption above, we have  $\log(Ng_{si}(0)) - \log(Ng_{sj}(0)) = c_{si} - c_{sj} = -FD * \log(\epsilon_{si}/\epsilon_{sj})$ , thus  $(a_{si} - a_{sj})\theta = (b_{si} - b_{sj})$ , the right part of the equation is a constant, while  $\theta$  is a variable in  $(-\infty, +\infty)$ , hence  $a_{si} - a_{sj} = 0$ ,  $b_{si} - b_{sj} = 0$ . Therefore, the fitted coefficients  $\mathbf{a} = a_{s1} = a_{s2} = \dots = a_{sS}$ ,  $\mathbf{b} = b_{s1} = b_{s2} = \dots = b_{sS}$ . The procedure of this proof can be easily reversed.  $\square$

Note that the value of the coefficients  $\mathbf{a}$ ,  $\mathbf{b}$  will change if use logarithms with different base. However, this only scales the range of them, while we aim at exploring the relative values. For instance, if we use the base 2 logarithm and the base 10 logarithm on  $Ng$  respectively, then  $\mathbf{a}_{\log_2} / \mathbf{a}_{\log_{10}} = \log_2(Ng) / \log_{10}(Ng) = \log_2(10)$ .

For each rotation axis there's a coefficient  $\mathbf{a}$ , i.e., we will get a vector  $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_K\}$  if  $K$  rotation axes are used. Since  $\log(Ng(\theta))$  has the maximal value when  $\theta = 0$ , theoretically  $\mathbf{b} = 0$ . Thus we report  $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_K\}$  as the shape representation, besides, denote by  $L$  the length of projection of the trajectory on the rotation axis, we report  $\{L_1, L_2, \dots, L_K\}$  to represent the trajectory area.

LEMMA 2. *Theoretically,  $\mathbf{a}$  is in the range of  $[\log_{base}(\cos(1)), 0]$  when  $base > 1$ , and in  $[0, \log_{base}(\cos(1))]$  when  $0 < base < 1$ .*

PROOF. For  $\log_{base}(Ng(\theta)) = a\theta^2 + b\theta + c$ , let  $\theta = 1$ , since  $\mathbf{b} = 0$ , we get  $\mathbf{a} = \log_{base}(Ng(1)/Ng(0))$ . If  $Ng(1) = Ng(0)$ ,  $\mathbf{a} = 0$ , this happens when a trajectory equals all its projective trajectories, for instance, a line segment rotates around its  $MA$ . Another bound of  $\mathbf{a}$  is  $\log_{base}(\cos(1))$ , this happens in the cases such as a line segment or the Hilbert curve rotate round their  $MI$ . For a line segment,  $Ng(1)/Ng(0) \approx Length(\theta = 1)/Length(\theta = 0) = \cos(1)$ ; for the Hilbert curve,  $Ng(1)/Ng(0) \approx Area(\theta = 1)/Area(\theta = 0) = \cos(1)$ , where  $Length$  and  $Area$  denote the length of a projective line segment and the area that a projective Hilbert curve occupies respectively.  $\square$

Since we use the natural logarithm in this paper, if the  $\log(Ng(\theta)) - \theta$  curve is well fitted by the quadratic function, then  $\log(\cos(1)) \approx -0.6156 \leq \mathbf{a} \leq 0$ . However, according to our experiential knowledge, the minimal  $\mathbf{a}$  is between  $-0.8$  and  $-0.9$ . Though a difference exists between theoretical value and observed value, generally, our method effectively represents the shape of trajectories. If we use  $R^2$  (squared correlation coefficient) to evaluate the goodness-of-fit, all  $R^2$  values with respect to curves in Fig. 3(e)-3(h) are above 0.9.

## 4. PROPOSED ALGORITHM

Algorithm 1 outlines the key process of modeling vessel trajectory shape. By rotating a trajectory around each rotation axis, it extracts the topological feature of the trajectory with respect to each rotation axis, while  $FD$  only captures the feature that a trajectory whether full in space.

LEMMA 3. *The computation time for the coefficients is linear on the number of points  $N$  of a trajectory and linear on the grid number  $m$ .*

PROOF. The time of calculation of the geometry center  $O$ , finding the  $MA$ , and coordinates transformation are all  $O(N)$ . Given a grid number  $m$ , to compute  $Ng$ , we only go over the

---

### Algorithm 1: Vessel Trajectory Shape

---

**Input:** Trajectory  $Tr$ ; rotation angles with ascending order  $\{\theta_1, \theta_2, \dots, \theta_j\}$ ,  $\theta_j \in [-\pi/2, \pi/2]$ ; grids number  $m$ ; axes number  $K$ .  
**Output:** Coefficients  $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_K\}$ .  
**1:** Get geometric center  $O$  of  $Tr$ , find the major axis  $MA$  and determine other  $K - 1$  rotation axes;  
**2:** Use  $MA$  and  $MI$  as the vertical axis and the horizontal axis,  $O$  as the origin, put  $Tr$  into the new coordinates system, and partition the new  $L_{MA} * L_{MI}$  space into  $m$  grids;  
**3:** Rotate  $Tr$  around each rotation axis according to the rotation angles  $\theta_j$ , and compute  $J$  projective trajectories for each rotation axis;  
**4:** Compute non-empty grids  $Ng$  for each projective trajectory;  
**5:** Fit a quadratic function for each rotation axis, report  $\mathbf{a}_k$ .

---

points of each projective trajectory a few times, and go over the grids once. The time of curve fitting is  $O(1)$ . So the total computation time is  $O(N)$  on  $N$  and  $O(m)$  on  $m$ .  $\square$

## 5. EXPERIMENTS

All experiments were run in the MATLAB R2013b 64-bit program on a PC with Intel Core i7-4790 CPU at 3.60 GHz, 32-GB RAM equipped with Windows 7. Default settings: trajectory points  $N = 5000$  (linear interpolation was utilised if  $N$  was less than that); rotation angles:  $\{\theta_1, \theta_2, \dots, \theta_{18}\}$ ,  $\theta_j \in [-\pi/2, \pi/2]$ ; grid number  $m = \{2^{10}, 2^{12}, 2^{14}\}$  and we took the mean of the results. Here we consider two rotation axes:  $MA$  and  $MI$ , thus we report  $\{\mathbf{a}_{MA}, \mathbf{a}_{MI}\}$  and  $\{L_{MA}, L_{MI}\}$  for each trajectory.

**Dataset. Synthetic data.** (i) Polygons (square, rectangle, equilateral triangle) and circle. (ii) Sine function  $y = \sin(x)$ ,  $x \in [0, 16\pi]$  and spiral curve  $x = t * \cos(20\pi t)$ ,  $y = t * \sin(20\pi t)$ ,  $t \in [0, 1]$ . (iii) Line segment and overlapping points. (iv) Hilbert curve and Cantor dust. (v) Random points,  $x \in [0, 1]$  and  $y \in [0, 0.1]$ . These datasets with different shapes were used to analyze and interpret the value of  $\mathbf{a}_{MA}$  and  $\mathbf{a}_{MI}$ , and test the computational efficiency.

**Real AIS data.** Our case study is based on two real-world AIS datasets: Shanghai Port and Xiamen Bay. Available information contains MMSI (Maritime Mobile Service Identity), longitude, latitude, speed, course, UTC time, ship name, ship type, ship length, ship width and draught. Details are shown in Table 1.

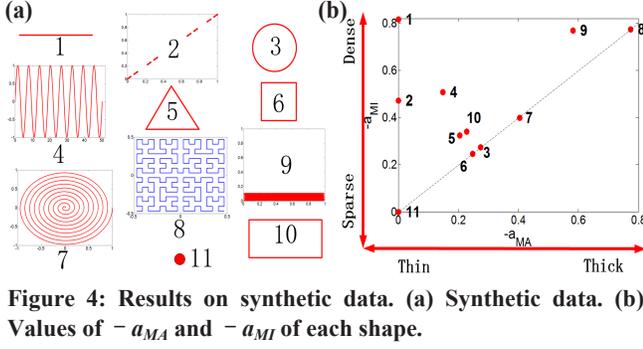
**Results on Synthetic Data. Results.** Fig. 4 presents the synthetic data and results. As the projection of trajectories on  $MA$  has the longest length, all points in Fig. 4(b) are above the diagonal. The way to interpret the  $MA$ - $MI$  Graph is as follows:

- (1) From left to right, shape varies from line-like to plane-like.
- (2) From bottom to up, the space a trajectory occupies varies from sparse to dense.

(3) Point lies on the diagonal indicates that  $L_{MA} = L_{MI}$ .

(4) Points of polygons and circle locate similarly in the graph. Points of 'dense plane' (datasets 8 and 9) lie on the right up corner and line-like (datasets 1 and 2) shapes lie on the left side.  $\mathbf{a}_{MA}$  and  $\mathbf{a}_{MI}$  of overlapping points dataset are both 0. For sine function and spiral curve, when increase  $x$  and  $t$  respectively, sine function will behavior like a line so its point in  $MA$ - $MI$  Graph will close to the left side, and the point of spiral curve will close to the right side and move along the diagonal.

**Efficiency analysis.** Table 2 shows the average running time on synthetic data under different  $N$  and  $m$ . We first fixed grid number



**Figure 4: Results on synthetic data. (a) Synthetic data. (b) Values of  $-a_{MA}$  and  $-a_{MI}$  of each shape.**

$m$  to  $2^{10}$  and varied  $N$  from 0.1 million to 0.5 million, then we fixed  $N = 5000$ , and varied  $m$  from  $2^{18}$  to  $2^{26}$  to examine the running time. The results show that the running time is linear on the dataset size and grid number in practice.

*The Goodness-of-fit.* It evaluates if the  $\log(Ng(\theta))-\theta$  curves are well fitted by the quadratic functions. Table 3 shows the result by using  $R^2$  as criteria,  $N = 5000$ ,  $m = 2^{10}$ , and the rotation axis is  $MI$ . Note that  $a = 0$  for the overlapping points (dataset 11), so there's no quadratic function. All curves are well fitted except the cases of square (dataset 6) and rectangle (dataset 10).

**Discovery of Vessel Patterns.** Vessel Trajectories were processed into moves and stops according to a speed threshold (1 knot). Scan all points by time order, a consecutive points set with speed  $\geq 1$  knot were considered as a move and those with speed  $< 1$  knot were considered as stops.

*Shapes vs. Speed.* Fig. 5(a)-5(d) show the results on stops and moves of Xiamen and Shanghai respectively. Low speed trajectories tend to behavior more like planes, while moves close to line segment. Generally, the stops have bigger values of  $-a_{MA}$ .

*Anomaly detection.* In Fig. 5(b), two points in blue circle are outliers compared to others, and have a relative bigger  $-a_{MA}$ , which means the two move trajectories follow more complicated shapes than other move trajectories. In Fig. 5(e), the trajectories look very strange and each speed record  $> 8$  knots. Maybe the ships were engaging in operations related to the port.

Fig. 5(d) shows  $-a_{MA}$ ,  $-a_{MI}$  and  $L_{MA}$  of Xiamen stops. It's interesting that there're two line-like trajectories with obviously bigger  $L_{MA}$  values, this means they moved a large range on the map with very low speed. In Fig. 5(f), all speed of their points  $< 0.1$  knots. Since the two trajectories crossed a large distance during a relative short time period, this is impossible and may be caused by the mechanical failure of AIS system.

## 6. CONCLUSIONS

In this paper, a new method was proposed for transforming the topological feature of trajectory into vectors for shape representation. This is the first work to examine and provide theoretical analysis on the self-similarity characteristic of vessel trajectories. It effectively modeling trajectory shape and is efficient enough to be performed in practical applications. As many trajectory datasets show the property of self-similar, it may be easily extended to closed environments, such as finding patterns from car or human trajectories.

**Table 1. Details of the two AIS datasets.**

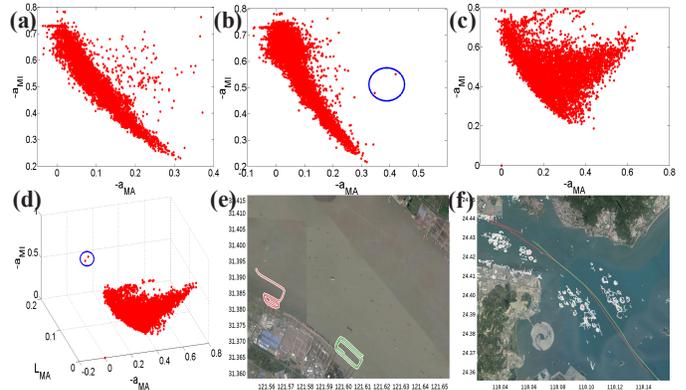
Dataset	Area	Time	Points	Ships
Shanghai	Lon: 121.105	2011.10.26	1,190,856	6,526
	~122.5 ° E	00:00:00 ~		
	Lat: 30.61	2011.10.31		
Xiamen	~31.885 ° N	23:59:59		
	Lon: 117.724	2015.09.03	613,315	1,216
	~118.383 ° E	00:00:00 ~		
	Lat: 24.350	2015.09.12		
	~24.710 ° N	23:59:59		

**Table 2. Running time vs.  $N$  and  $m$ .**

$N(m=2^{10})$	0.1mil	0.2mil	0.3mil	0.4mil	0.5mil
Time	0.483s	0.999s	1.486s	1.985s	2.474s
$m(N=5000)$	$2^{18}$	$2^{20}$	$2^{22}$	$2^{24}$	$2^{26}$
Time	0.287s	1.184s	8.904s	41.110s	176.70s

**Table 3. Goodness-of-fit  $R^2$ .**

Dataset	1	2	3	4	5
$R^2$	0.964	0.930	0.998	0.993	0.940
Dataset	6	7	8	9	10
$R^2$	0.399	0.999	0.972	0.937	0.602



**Figure 5: Results on stops and moves. (a) Xiamen moves. (b) Shanghai moves. (c) Shanghai stops. (d)  $-a_{MA}$ ,  $-a_{MI}$  and  $L_{MA}$  of Xiamen stops. (e)-(f) Detected anomalies.**

## 7. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grants No. 71571186, 61273322, 71471176 and 61703416).

## 8. REFERENCES

- [1] Harati-Mokhtari, A., Wall, A., Brooks, P. and Wang, J., "Automatic Identification System (AIS): data reliability and human error implications," *The Journal of Navigation*, vol. 3, no. 60, pp. 373-389, 2007.
- [2] Li Y., Li Y., Gunopulos D., and Guibas L., "Knowledge-based trajectory completion from sparse GPS samples," in *ACM SIGSPATIAL '16*, 2016.
- [3] Thakur G. S., Hui P. and Helmy A., "Evidence of long range dependence and self-similarity in urban traffic systems," in *ACM SIGSPATIAL '15*, 2015.
- [4] Belussi A. and Faloutsos C., "Estimating the selectivity of spatial queries using the 'correlation' fractal dimension," in *VLDB '95*, 1995.
- [5] Matsubara Y., Li L., Papalexakis E., et al., "F-trail: finding patterns in taxi trajectories," in *PAKDD '13*, 2013.
- [6] Schroeder, M. R., "Fractals, chaos, power laws: minutes from an infinite paradise," W.H. Freeman and Company, New York, 1991.