applied sciences

MDPI

Article

Vessel Trajectory Reconstruction Based on Functional Data Analysis Using Automatic Identification System Data

Myeong-Hun Jeong [1], Seung-Bae Jeon [1], Tae-Young Lee [1], Min Kyo Youm [2] and Dong-Ha Lee [3,*]

[1] Department of Civil Engineering, Chosun University, Gwangju 61452, Korea; mhjeong@chosun.ac.kr (M.-H.J.); zeon6779@gmail.com (S.-B.J.); taeyounglee1@naver.com (T.-Y.L.)
[2] Center for Built Environment (CBE), Sungkyunkwan University, Suwon-si 16419, Gyeonggi-do, Korea; tomsmith850918@gmail.com
[3] Department of Civil Engineering, Kangwon National University, Chuncheon-si 24341, Korea
* Correspondence: geodesy@kangwon.ac.kr

Received: 24 December 2019; Accepted: 21 January 2020; Published: 28 January 2020

Abstract: This study provides an automatic shipping-route construction method using functional data analysis (FDA), which analyzes information about curves, such as multiple data points over time. The proposed approach includes two steps: outlier detection and shipping-route construction. This study uses automatic-identification system (AIS) data for the experiments. The effectiveness of the proposed method is demonstrated through case studies, wherein our approach is compared with the Mahalanobis distance method for trajectory-outlier detection, and the performance of vessel trajectory reconstruction is compared with that of a density-based approach. The proposed method improves understanding of vessel-movement dynamics, thereby improving maritime monitoring and security.

Keywords: map construction; shipping-route construction; functional data analysis; data depth

1. Introduction

Advances in location-aware technology are resulting in the generation of massive trajectory data. A large volume of movement data enhances the determination of interesting and hidden patterns from big spatiotemporal data in various domains. Map construction is a data-mining technique that automatically produces street maps from trajectory data [1]. This paper proposes a new method that constructs maps (i.e., shipping routes) from vessel-movement data.

Street-map construction traditionally requires costly aerial surveying or labor-intensive field surveying. With the advent of GPS-enabled technology, volunteered geographic information (VGI), such as OpenStreetMap (OSM), have provided a valuable source of map datasets [2]. However, only authorized users can update maps in OSM. Thus, it is necessary to develop automatic map-construction algorithms from moving-object data.

While extensive research has been carried out on the construction of street maps from tracking data [3–5], there are limited studies regarding the construction of shipping routes. For example, the Korea Hydrographic and Oceanographic Agency (KHOA) not only conducts hydrographic surveys, but also provides navigational information to prevent marine accidents. Although KHOA provides nautical charts, they do not provide shipping routes. Thus, shipowners sail ships using charts that present the water depth and land height.

Therefore, it is essential to construct shipping routes for maritime safety. This study utilizes autonomous identification system (AIS) data, which is a self-reporting maritime system. The AIS

equipment continuously reports vessel information, ranging from vessel heading course, speed, and position to maritime mobile satellite identity (MMSI) [6]. This study applies functional data analysis (FDA) to construct shipping routes from the AIS data. FDA is a method of understanding multiple data points over time as a single block of a curve [7]. Classical statistics is the statistical analysis of measurements made up of single numbers. However, in multivariate statistics, different values are derived for each subject and unit. As data collection by various methods becomes complicated, it is reasonable to observe the collected data in one continuum. In short, the aim of this study is to generate shipping routes from AIS data using FDA, which strengthens maritime safety.

Following a review of related work (Section 2), Section 3 precisely presents the proposed map-construction method based on FDA. Section 4 then presents an experimental evaluation and comparisons of the proposed method. Finally, the paper concludes with a discussion of the limitations of the approach, and future work (Section 5).

2. Background

Map construction is the process of generating street maps from moving objects' traces, such as GPS coordinates. The final output is usually represented as an embedded graph. Map-construction algorithms can be categorized into point clustering, incremental track insertion, and intersection linking [1].

Point clustering regards trajectories as sets of points, and generates maps by clustering these points [8]. A K-means clustering algorithm is employed to cluster the input points [9]. The centers of the clusters are connected after the clusters have been refined. An image-processing technique is applied to generate a map [10]. The total length of the trajectory is calculated for each cell across the rectangular grid for all map areas. Then, Gaussian blur is applied to remove small gaps caused by GPS noise. Next, the polygon boundaries of the road area are extracted, and a road center-line is created. However, this algorithm is density based and its accuracy depends on the size of the grid cells created on the map. Furthermore, outliers in the datasets have a negative effect on the generation of maps in density-based algorithms.

The incremental track-insertion algorithm inserts tracks into an empty graph based on map-matching ideas [11]. The incremental construction of the map, such as the addition or deletion of tracks, is based on distance measurements and vehicle headings [1]. For example, the previous research [4] proposes two steps in the map-construction algorithm. First, the algorithm identifies a portion of the trajectory that exists on the currently partitioned map. This partitioned map is generated using the Frechet-distance transformation method. The next step inserts inconsistencies into the current map by adding new boundary points and borders as needed. The algorithm requires only the time and δ parameters to split the trajectory, where the δ parameter is the minimum distance between roads and the minimum distance between two intersections. This algorithm has significant advantages when the trajectory added at all stages is correct, but the disadvantage of adding the wrong trajectory has a negative effect on the overall results. Similarly, study [12] proposes a map-construction algorithm that completes the graph in two steps: preprocessing and clarification. The preprocessing step reduces the straight part of the oversampled trajectory while dividing the wrong trajectory and preserving the sample of the high curvature. The clarification step decreases the effect of noise on the input trajectory by using a forced algorithm to clean the initial data.

Lastly, intersection-linking algorithms first detect intersection vertices and then construct edges using these vertices [1]. Intersections are identified based on movement characteristics (speed, direction) or point density. For instance, study [13] identifies potential turning points according to changes in the direction and speed. Next, the cross node obtained from the input trajectory is connected to the edge. This algorithm is highly useful when the vertex is part of an intersection.

These aforementioned methods predominantly focus on vehicle data to generate street maps. This study deals with ship movement data. Moreover, FDA has not yet been applied to construct maps. Our approach utilizes FDA to not only remove outliers from ship trajectories, but also to produce shipping routes. The following section explains the details of the proposed method.

3. Methods

FDA aims to analyze information regarding curves. It has been used to examine temperature variation, gene expression signals, and human-movement comparisons [14]. The advantage of the FDA is that it derives significant sources of pattern and variation between the collected data. It is commonly used to detect outlying curves, based on a functional boxplot by [15]. This method is based on the notion of the modified band depth [16]. To understand the modified band depth, it is necessary to comprehend the concept of data depth.

Data depth is a robust statistic with which to measure the centrality of the underlying data. Data depth orders data by their degree of centrality [17,18]. A variety of data depth functions has been proposed: simplicial depth, projection depth, and Mahalanobis depth [19–21].

The simplicial depth partitions the underlying data into a set of $\begin{pmatrix} n \\ d+1 \end{pmatrix}$ unique $(d+1)$-simplices. Figure 1 illustrates the concept of the simplicial depth. A subset of the $\begin{pmatrix} n \\ 3 \end{pmatrix}$ 3-simplicies (triangles) in $\mathbb{R}^2$ is defined by a set of objects $(x_1, x_2, x_3) \in X$, where $x \in \mathbb{R}^d$ regarding the underlying data $X = \{x_1, \ldots, x_n\}$. The point triangle in Figure 1 is regarded as the deepest point within the underlying data because many simplices (triangles) contain it. However, the square symbol represents the smallest depth in the underlying data. Any simplices (triangles) that three points make do not contain it. Simplices depth presents the centrality of each point by describing how much each point lies in the cloud of points.

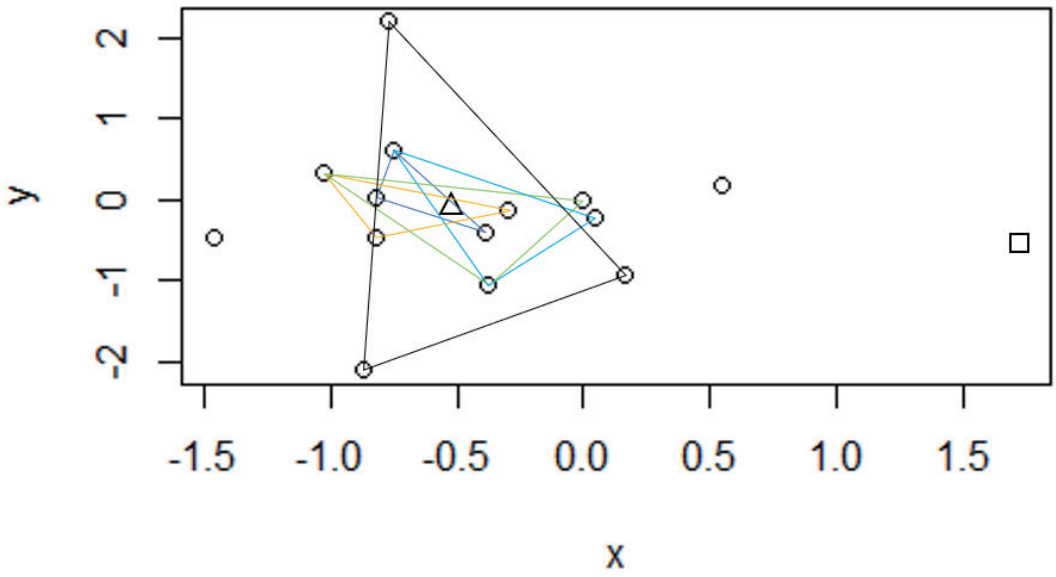

Figure 1. Simplicial data depth.

The band depth extends the concept of data depth for FDA. Let a function $y(t)$ be the subset of the plane $G(y) = \{(t, y(t)) : t \in T\}$, where t is time or a value of the function. The band is the subset of the plane that lies between the two curves y_{i1} and y_{i2} such that $BD(y) = \{(t, x) : t \in T, min(y_{i1}(t), y_{i2}(t)) \leq x \leq max(y_{i1}(t), y_{i2}(t))\}$ [22]. If the band depth of y is high, y is deeply nested by other curves.

While the band depth does not take into account the proportion of time that a curve $y(t)$ lies between $y_{i1}(t)$ and $y_{i2}(t)$, the modified band depth considers the proportion of time that $y(t)$ contains a value in the band [16]. For example, there are six possible bands determined by two curves in Figure 2. $BD(y_4) = 3/6 = 0.5$. $y(4)$ is only contained in the bands delimited by itself and other curves (i.e., $y(1)$ and $y(4)$, $y(2)$ and $y(4)$, and $y(3)$ and $y(4)$). However, let us consider the modified band

depth, $MBD(y_4) = (3 + 0.5 + 0.5)/6 = 0.67$. The modified band depth considers the proportion of times that the curve y_4 in the band. For instance, the modified band depth of y_4 in the band delimited by y_1 and y_3 is 0.5. Similarly the modified band depth of y_4 in the band delimited by y_2 and y_3 is also 0.5.

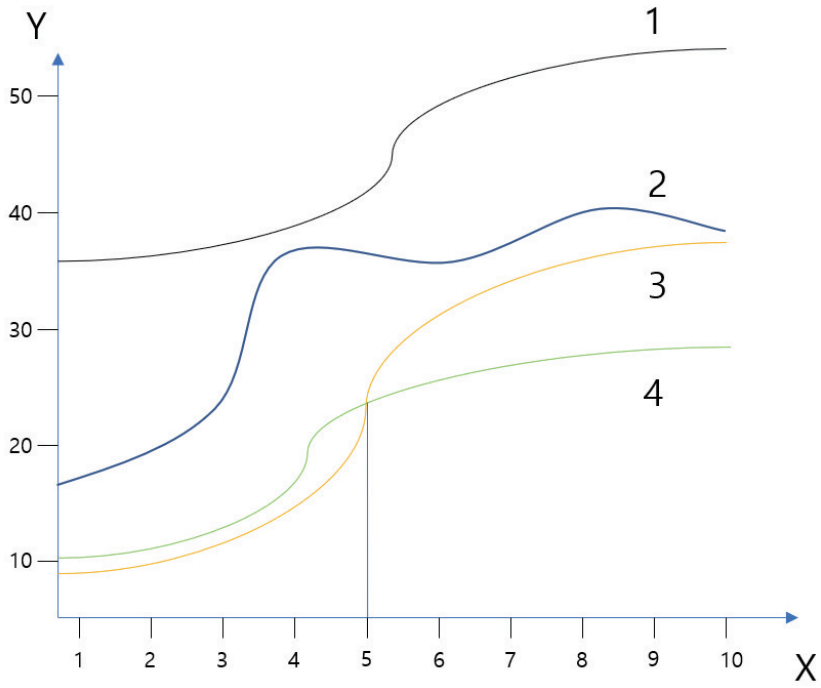

Figure 2. Modified band depth.

To apply FDA to AIS data, we anchor one of the directions. For example, Figure 3 represents vessels moving in the north and south directions. If we consider the north and south direction as the basis of FDA, such as $y(t)$, the values of $y(t)$ is the mean of the east-west locations. This study uses a robust statistic, such as trim mean (20%), to calculate the variation in the east-west locations of vessels. In terms of the interval of the north-south axis, this study uses 500 m owing to the granularity of the AIS data.

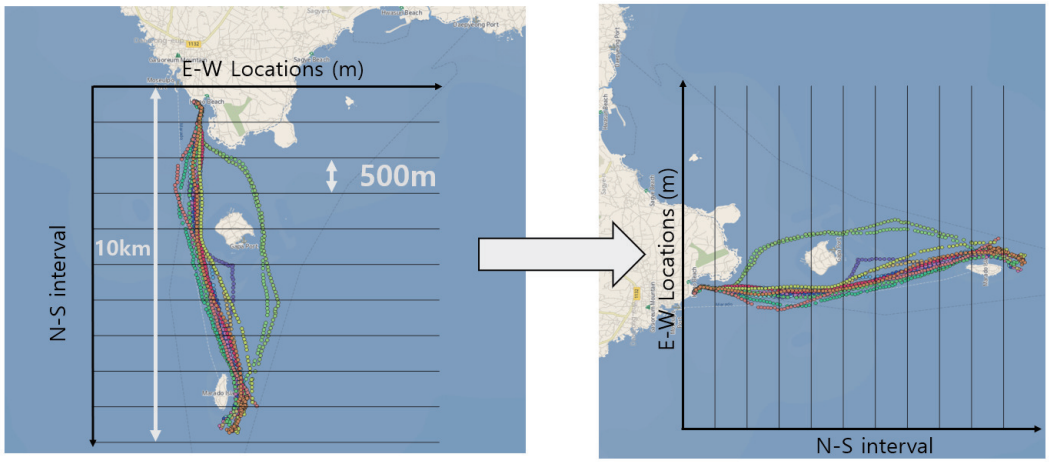

Figure 3. Functional data analysis for shipping routes.

Then, a functional boxplot [15] is applied to detect outliers from the vessel trajectories, and the identified vessel trajectories are removed. Finally, we describe the shipping routes based on the centerline from the function data analysis.

4. Experiments

4.1. Data

This study used AIS data collected in 2014, from the Republic of Korea. The data contains over 240 million ship-trip records (130 GB). Two datasets were generated for the experiments. One is the ship trajectories from Jeju Island Unjin port to Mara Island Saledok port over three days (Case 1). The other is the ship movements from Jeju Island Sagye port to Mara Island Saledok port over three days (Case 2). Jeju island is located in the south of the Republic of Korea and Mara Island is in the southernmost Korea. Figures 4 and 5 present the two experimental AIS datasets for this study.

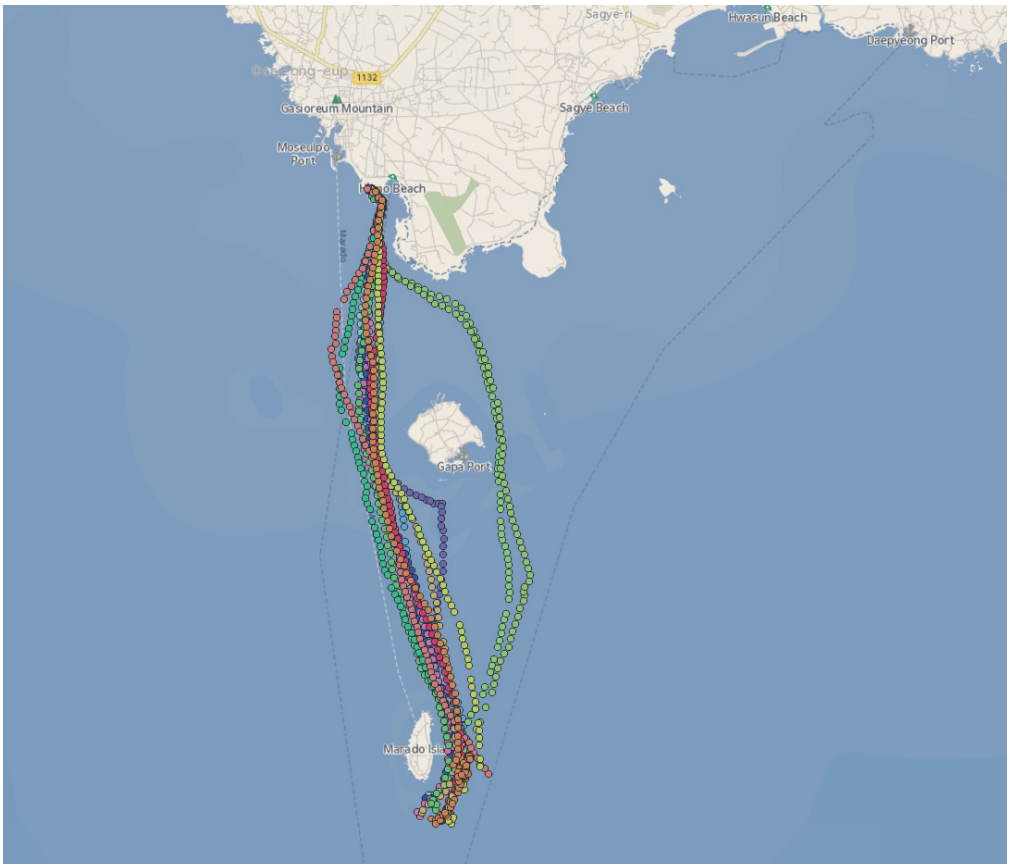

Figure 4. Case 1 dataset.

Furthermore, this study used Amazon Web Service (AWS) to deal with the large volume of vessel-movement data. Data was uploaded into an Amazon S3 bucket. Then, we used Hadoop with Pig (i.e., Elastic MapReduce in AWS) for the data-cleaning process. This study used R to implement the functional data analysis. Our functional data analysis is based on the R package FDA [23]. Figure 6 presents the system architecture used in this study.

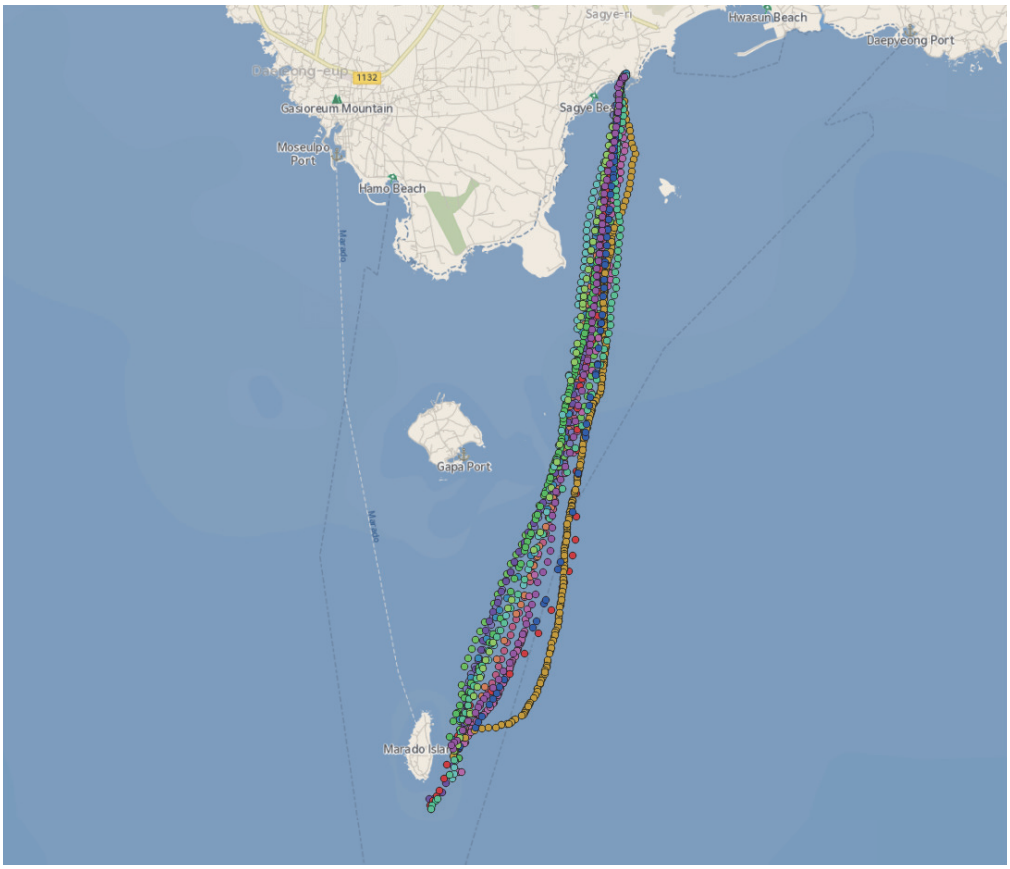

Figure 5. Case 2 dataset.

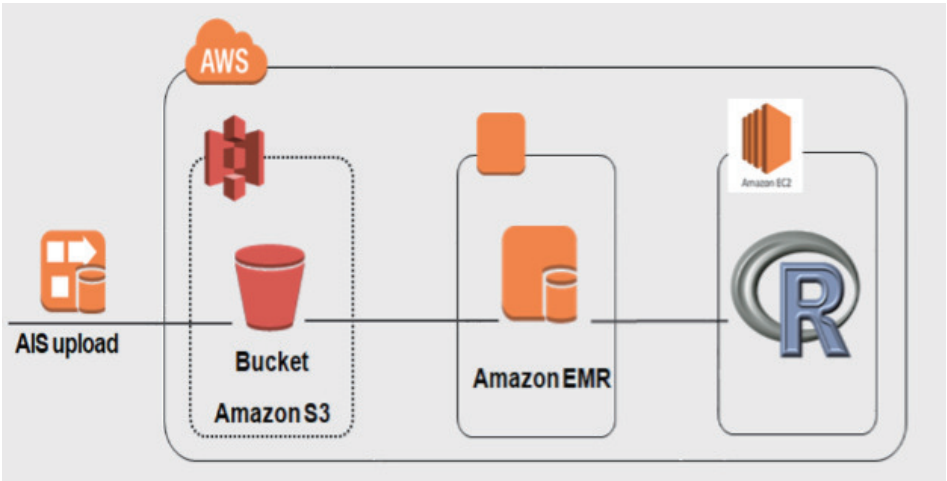

Figure 6. System architecture.

4.2. Case 1: Jeju Island Unjin Port to Mara Island Saledok Port

There are thirteen trajectories in the Case 1 dataset. We first detected outliers from the trajectories. The functional boxplot in Figure 7 presents the box, whiskers, median curve, and two outliers. The magenta box indicates the 50% central region, which is the region through which most trajectories pass. The whiskers with blue curves denote envelopes, which provide the 25% and 75% central regions. The red dashed lines indicate outliers, which are identified as 1.5 times the 50% central region [15].

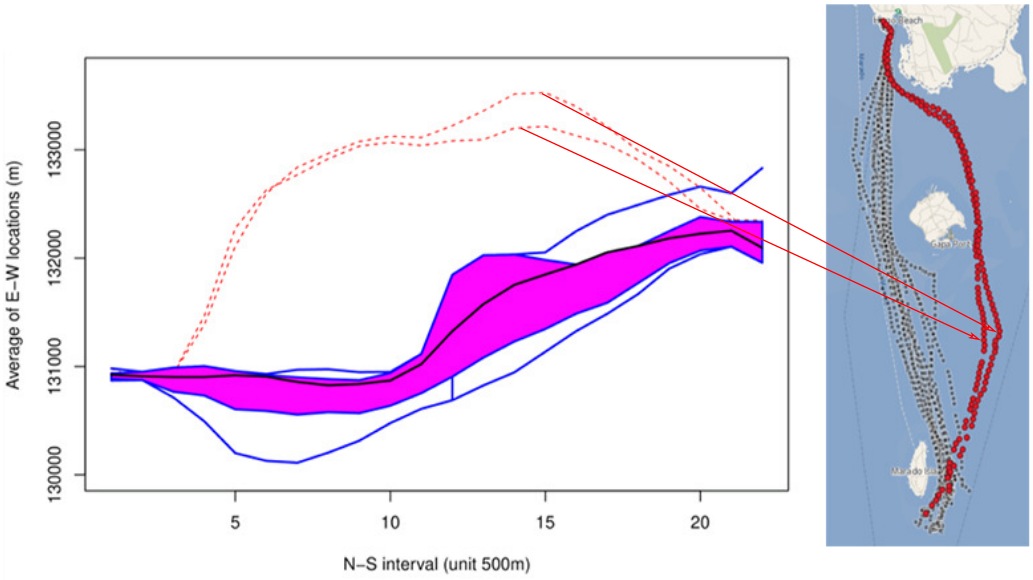

Figure 7. Outlier detection using functional boxplot.

Figure 7 shows the two detected outliers with comparatively non-typical patterns. We divided the north-south axis into 500 m increments, as indicated by the *x*-axis in Figure 7. The *y*-axis shows the average longitude of the vessel locations along the east-west axis per each interval. The two identified outliers are represented on a map to the right of the graph.

While the functional boxplot identified two trajectories as outliers, the Mahalanobis distance method detected one outlier in Figure 8. The Mahalanobis distance method is a multivariate outlier-detection method [14]. This study used two parameters (i.e., distance and time) for the Mahalanobis distance method. Interestingly, Figure 8 shows that one trajectory had a considerably longer time compared with the other trajectories, although this trajectory is one of the main patterns.

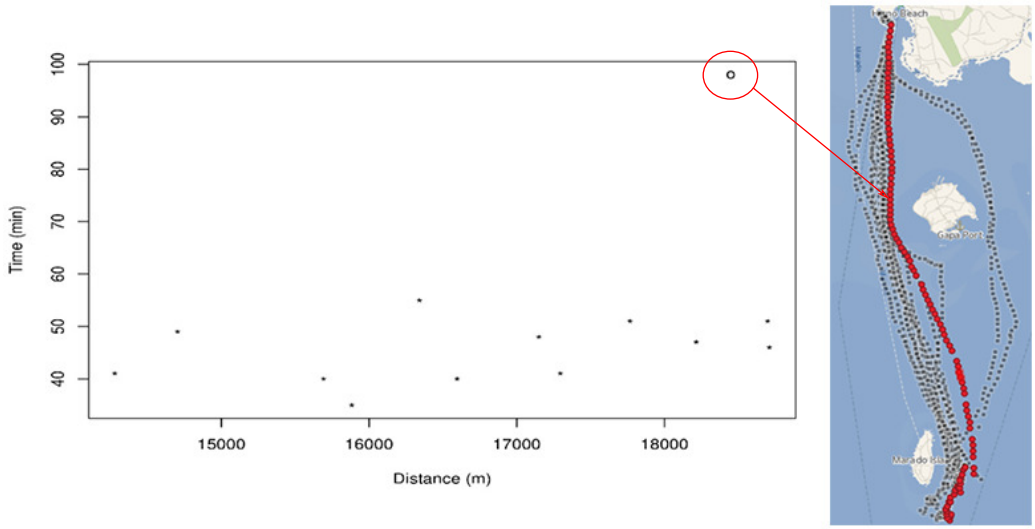

Figure 8. Outlier detection using Mahalanobis distance method.

A functional boxplot tends to be more concerned with trajectory patterns without explicit time information, whereas the Mahalanobis distance method suffers from identifying different trajectory patterns. It is worth noting that although it is easy and fast to calculate the Mahalanobis distance, this method is not robust because the high correlation of variables makes harder the calculation of the inverse of the correlation matrix in the Mahalanobis distance [24].

In constructing the shipping routes, the previously identified outliers were removed, and the median curve was generated. The final shipping routes generated by the proposed method are shown in Figure 9a. Further, this study used an alternative method [10] to generate shipping routes. The algorithm in [10] uses a density-based technique to construct the map by varying the cell sizes. This study changed the cell sizes ranging from 119 m × 380 m to 238 m × 760 m in Figures 9b–e.

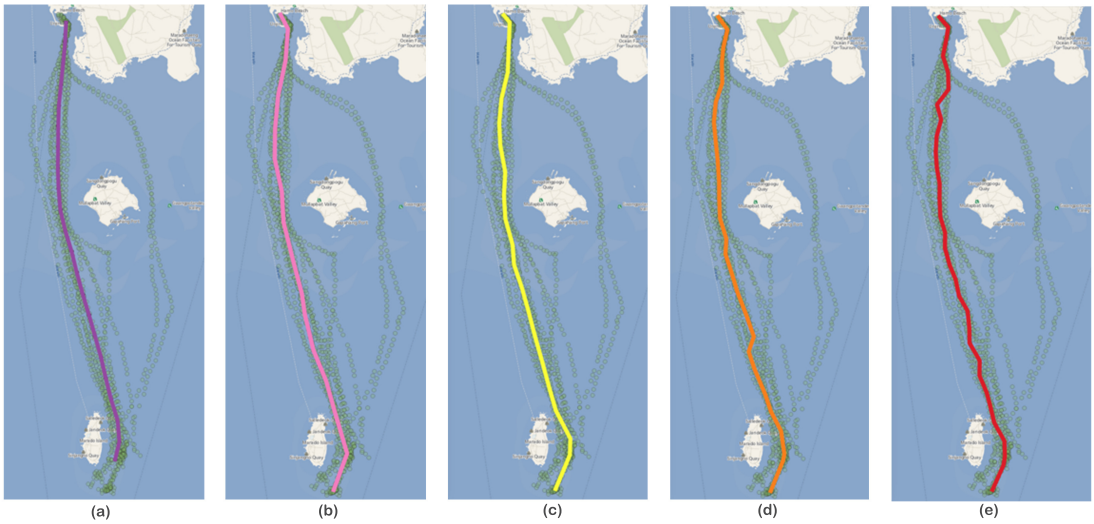

Figure 9. Shipping routes construction: (**a**) functional data analysis; (**b**) density-based technique (cell size: 238 m × 760 m); (**c**) density-based technique (cell size: 179 m × 571 m); (**d**) density-based technique (cell size: 143 m × 456 m); (**e**) density-based technique (cell size: 119 m × 380 m).

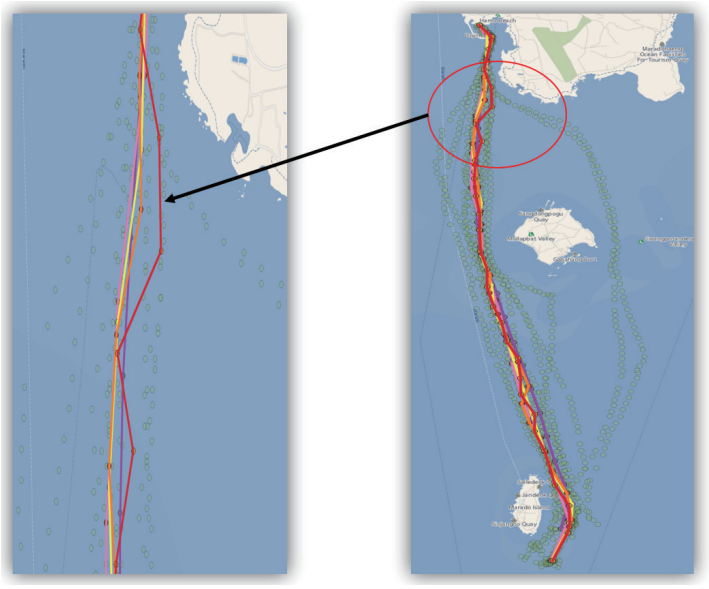

Figure 10. Comparison of all shipping-routes constructions in Case 1.

In addition, all results of shipping routes construction were superimposed on the same map in Figure 10. In particular, Figure 10 clearly shows that the density-based approach is highly affected by cell size. Discretization in a regular grid generated irregular shipping routes. The following section describes another dataset that was evaluated using the same method.

4.3. Case 2: Jeju Island Sagye Port to Mara Island Saledok Port

This section describes the shipping-route construction using a different dataset. There are fifteen trajectories in the Case 2 dataset. In a similar vein, the proposed method detected two anomalous trajectories, and the detected outliers are represented on the right-side map in Figure 11.

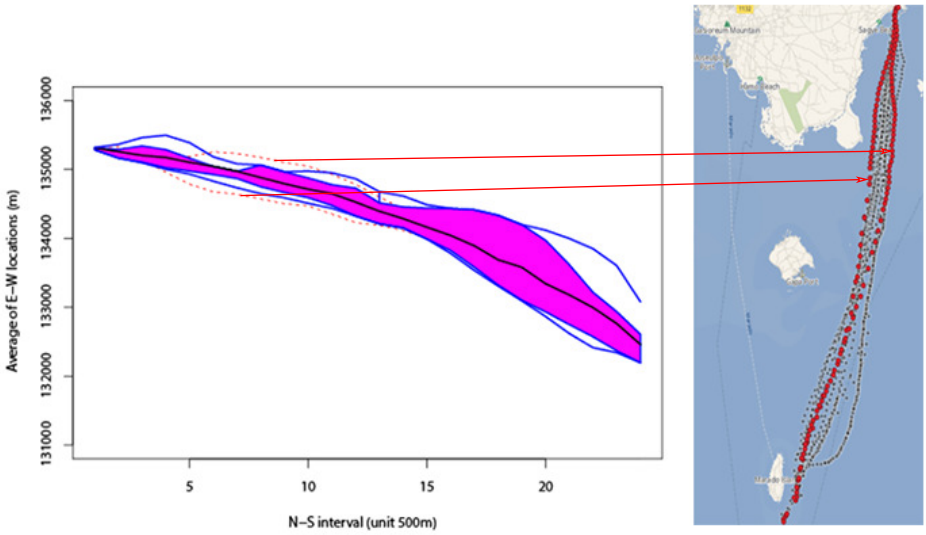

Figure 11. Outlier detection using functional boxplot.

The shipping routes were generated using the median curve from the functional data analysis in Figure 12a. Further, a density-based technique produced shipping routes by varying the cell sizes. Case 2 of this study changed the cell sizes ranging from 124 m × 440 m to 248 m × 879 m in Figure 12b–e.

Additionally, all results were superimposed on the same map in Figure 13. Overall, this case supports the previous view that the selection of appropriate cell size is vital to a density-based method.

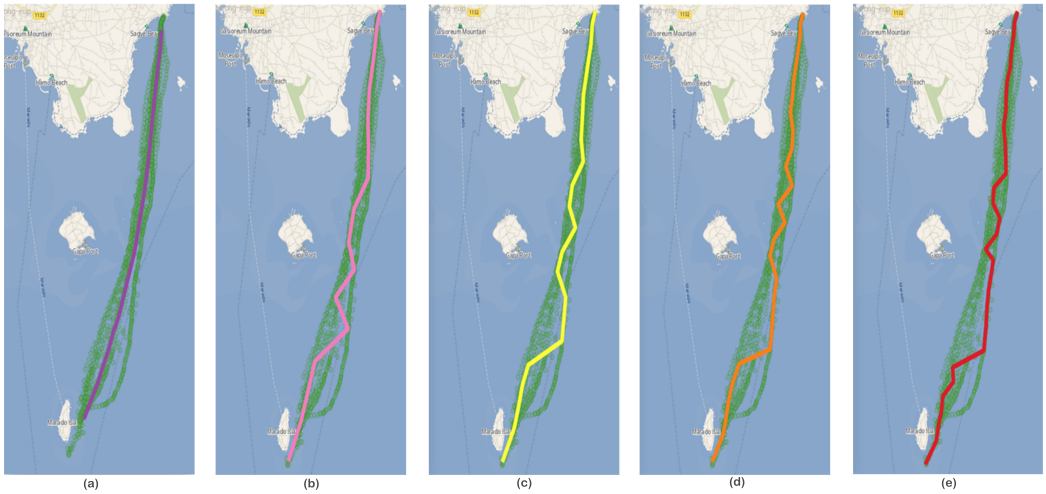

Figure 12. Shipping routes construction: (**a**) functional data analysis; (**b**) density-based technique (cell size: 248m × 879m); (**c**) density-based technique (cell size: 186 m × 660 m); (**d**) density-based technique (cell size: 149 m × 528 m); (**e**) density-based technique (cell size: 124 m × 440 m).

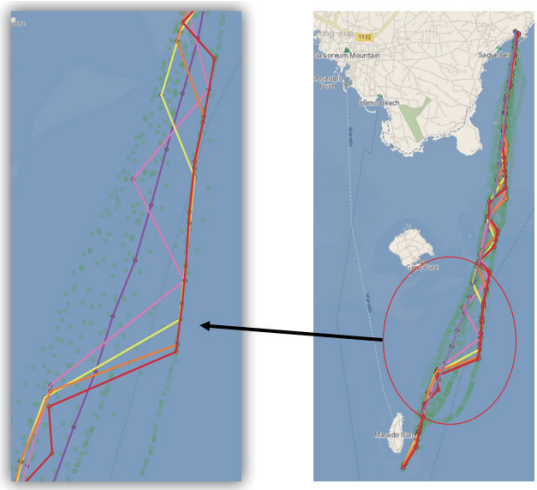

Figure 13. Comparison of all shipping-routes constructions in Case 2.

5. Discussion and Conclusions

This study describes a new method for the construction of shipping routes using the FDA. The proposed method starts by detecting anomalous trajectories and removing them. Then, the median curve from the functional data analysis is identified; this is a shipping route.

Regarding the detection of trajectory outliers, this study compared the functional data analysis approach with the Mahalanobis distance method, which is a multivariate outlier detection method. FDA detects abnormal movement patterns. In contrast, the Mahalanobis distance method cannot detect them. It is necessary to look at the pros and cons of both methods, based on a dedicated application.

Furthermore, the resulting shipping-route constructions are compared with those of a density-based approach. It is significant that the grid cell size resolves the resolution of the map construction because small grid cell sizes cause irregular shipping routes that are not naturally present.

However, it is worth noting that the proposed method cannot be applied to all types of map construction, such as street-map construction. For example, it is unlikely that FDA can accurately represent all street maps at intersections. The proposed method is good at representing simple trajectories, such as shipping routes in the ocean.

Ultimately, further investigation should focus on the design of a robust algorithm for a variety of map-construction applications based on the FDA. Besides, the current research does not consider locational circumstances or natural environment. Such geographic contexts should be incorporated. This research will lead to increased knowledge and a better understanding of vessel movement dynamics, which improves maritime monitoring and security.

Author Contributions: Conceptualization, M.-H.J.; methodology, M.-H.J. and S.-B.J.; software, M.-H.J., S.-B.J. and T.-Y.L.; validation, M.-H.J. and S.-B.J.; formal analysis, M.-H.J., S.-B.J. and T.-Y.L.; investigation, M.-H.J., S.-B.J. and T.-Y.L.; resources, M.-H.J., S.-B.J. and T.-Y.L.; data curation, S.-B.J. and T.-Y.L.; writing—original draft preparation, M.-H.J.; writing—review and editing, M.-H.J. and M.K.Y.; visualization, M.-H.J. and S.-B.J.; supervision, D.-H.L.; project administration and funding acquisition, M.-H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2018R1C1B5043892).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ahmed, M.; Karagiorgou, S.; Pfoser, D.; Wenk, C. Map construction algorithms. In *Map Construction Algorithms*; Springer: Berlin, Germany, 2015; pp. 1–14.

2. Haklay, M.; Weber, P. Openstreetmap: User-generated street maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [CrossRef]
3. Agamennoni, G.; Nieto, J.I.; Nebot, E.M. Robust inference of principal road paths for intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.* **2010**, *12*, 298–308. [CrossRef]
4. Ahmed, M.; Wenk, C. Constructing street networks from GPS trajectories. In *European Symposium on Algorithms*; Springer: Berlin, Germany, 2012; pp. 60–71.
5. Biagioni, J.; Eriksson, J. Inferring road maps from global positioning system traces: Survey and comparative evaluation. *Transp. Res. Rec.* **2012**, *2291*, 61–71. [CrossRef]
6. Jeong, M.H.; Lee, D.H.; Lee, T.Y.; Lee, J.H. Robust local spatial autocorrelation analysis of massive vessel movements. *J. Coast. Res.* **2019**, *91*, 306–310. [CrossRef]
7. Ramsay, J.O. Functional data analysis. *Encycl. Stat. Sci.* **2004**, *4*. [CrossRef]
8. Duran, D.; Sacristán, V.; Silveira, R.I. Map construction algorithms: An evaluation through hiking data. In Proceedings of the 5th ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, Burlingame, CA, USA, 31 October 2016; pp. 74–83.
9. Edelkamp, S.; Schrödl, S. Route planning and map inference with global positioning traces. In *Computer Science in Perspective*; Springer: Berlin, Germany, 2003; pp. 128–151.
10. Davies, J.J.; Beresford, A.R.; Hopper, A. Scalable, distributed, real-time map generation. *IEEE Pervasive Comput.* **2006**, *5*, 47–54. [CrossRef]
11. Quddus, M.A.; Ochieng, W.Y.; Noland, R.B. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transp. Res. Part E Merg. Technol.* **2007**, *15*, 312–328. [CrossRef]
12. Cao, L.; Krumm, J. From GPS traces to a routable road map. In Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, Seattle, WA, USA, 3–6 November 2009; pp. 3–12.
13. Karagiorgou, S.; Pfoser, D. On vehicle tracking data-based road network generation. In Proceedings of the 20th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA, 6–9 November 2012; pp. 89–98.
14. Wilcox, R.R. *Introduction to Robust Estimation and Hypothesis Testing*, 4th ed.; Academic Press: Cambridge, MA, USA, 2016.
15. Sun, Y.; Genton, M.G. Functional boxplots. *J. Comput. Graph. Stat.* **2011**, *20*, 316–334. [CrossRef]
16. López-Pintado, S.; Jornsten, R. Functional analysis via extensions of the band depth. In *Lecture Notes—Monograph Series*; Institute of Mathematical Statistics: Shaker Heights, OH, USA, 2007; pp. 103–120.
17. Jeong, M.H.; Cai, Y.; Sullivan, C.J.; Wang, S. Data depth based clustering analysis. In Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, New York, NY, USA, 31 October–3 November 2016; p. 29.
18. Jeong, M.H.; Sullivan, C.J.; Gao, Y.; Wang, S. Robust abnormality detection methods for spatial search of radioactive materials. *Trans. GIS* **2019**, *23*, 860–877. [CrossRef]
19. Zuo, Y.; Serfling, R. General notions of statistical depth functions. *Ann. Stat.* **2000**, *28*, 461–482. [CrossRef]
20. Serfling, R. Depth functions in nonparametric multivariate inference. *DIMACS Ser. Discret. Math. Theor. Comput. Sci.* **2006**, *72*, 1.
21. Mosler, K., Robustness and Complex Data Structures. In *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather*; Chapter Depth Statistics; Becker, C., Fried, R., Kuhnt, S., Eds.; Springer: Berlin, Germany, 2013; pp. 17–34.
22. López-Pintado, S.; Romo, J. On the concept of depth for functional data. *J. Am. Stat. Assoc.* **2009**, *104*, 718–734. [CrossRef]
23. Ramsay, J.O.; Wickham, H.; Graves, S.; Hooker, G. FDA: Functional Data Analysis, 2018. R Package Version 2.4.8. Available online: https://cran.r-project.org/web/packages/fda/fda.pdf (accessed on 28 January 2020).
24. Varmuza, K.; Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics*; CRC Press: Boca Raton, FL, USA, 2016.