Kyay Mone Soe Oo
Shanghai Maritime University, China
Myanmar Maritime University, Myanmar
Chaojian Shi, Hu Qinyou
Shanghai Maritime University, China
Adam Weintrit
Gdynia Maritime University, Poland

CLUSTERING ANALYSIS AND IDENTIFICATION OF MARINE TRAFFIC CONGESTED ZONES AT WUSONGKOU, SHANGHAI

Shanghai, with its natural, cultural and historical wealth, is not only one of China's most beautiful cities, but it is also one of the most exciting cities in the world. However, there are enormous challenges for navigation in the Shanghai Strait due to its geographical, geopolitical and oceanographic structure. One of the challenges is the marine traffic which crosses from one side to other of the strait. In this study, an attempt is made to identify of vessel traffic zones based on DBSCAN in the Wusongkou. It is located along the north end of Huangpu river which flows from South-West of Shanghai to the North-East and flows into Yangtze river. Ship's domain is introduced into the DBSCAN algorithm, a particle suitable clustering algorithm is improved for clustering the real-time ship's dynamic data and detecting potential traffic congested areas at sea, and define three neighborhood models. In addition, fuzzy evaluation model is applied to identify traffic congestion degree. At the end of study, combining the improved DBSCAN algorithm and fuzzy evaluation model for traffic congestion degree, using three neighborhood models with different size to analyses the AIS data from the vessels nearby Wusongkou in Shanghai, and build the corresponding figure of traffic condition visualisation, used to visualise the evaluation result. The result indicate that the neighborhood three model (length is seventeen times of ship's length, width is six point four times of ship's length plus ship's width) can identify the traffic congested zones better.

INTRODUCTION

In recent years, shipping is developing rapidly over the world to meet the growing economic demands. Ships are getting greater, speedier, and more professional, and the number of ships improves dramatically. These factors make ports and straits more and more crowded and complicated, and the resulted traffic congestion or jam may enhance the risk of collision and decrease the traffic efficiency in a great extent. Therefore, the real-time identification of the congested zones at sea is vital role both on water traffic induction and its control.

The concept of marine traffic congestion degree and its calculation, however, have not been well developed like in the road transportation domain. In order to meet the requirements of collision avoidance and traffic management, more and more vessels have equipped AISs.

Clustering, groups database data into meaningful subclasses in such a way that minimizes the intra-differences and maximizes the inter-differences of these subclasses, is one of the most widely studied problems in data mining. Clustering technique is applied in many areas, such as statistical data analysis, pattern recognition, image processing, and other businesses applications.

The congested zones are usually detected by calculating the traffic flow and its density. Although intuitive, it has some shortages because it can not reflect the impact of traffic order and the vessels' dimension on the traffic congestion degree. In this paper, traffic velocity is used for detecting the congested zones.

This paper is organized as follow: Section 2 presents the features of marine traffic congestion, the traditional methods to determine it and their disadvantages. AIS (Automatic Identification System) is presented in section 3, Clustering algorithms are introduced in section 4, a summary on the DBSCAN Algorithm and Improvement are presented in section 5. In section 6, we performed fuzzy evaluation model to determine traffic congestion degree with main traffic flow velocity. An experimental evaluation of the effectiveness using DBSCAN is discussed in section 7. In Section 8, we performed Identification of congested zones. Finally, main conclusion is offered in Section 9.

Fig. 1. Wusongkuo in Shanghai [Google Maps]

1. MARINE TRAFFIC CONGESTION FEATURES AND TRADITIONAL DESCRIPTION OF TRAFFIC CONGESTION DEGREE

At present, there is no definite definition of traffic congestion degree of strait. In fact, marine traffic congestion always exists and can be manifested as:

- with low sailing velocity and speeding up and down frequently,
- with disorder navigation,
- with too many vessels blocked in the strait.

From the domestic and international research of marine traffic it is found that marine traffic engineers prefer to use traffic density or traffic volume to determine traffic congestion degree.

Traffic density is the instant average quantity of the vessels per unit area in the surveyed waters through a certain waters during a certain time period. Both traffic density and traffic volume cannot describe the above 1 and 2 features of marine traffic congestion. There are two other major disadvantages when traffic density and volume are applied to determine the marine traffic congestion degree. So traffic density or volume is not a perfect parameter to determine traffic congestion degree.

2. AUTOMATIC IDENTIFICATION SYSTEM (AIS)

2.1. AIS Data

Automatic Identification System (AIS) is a technology that was originally designed for collision avoidance, but it also can dramatically change the way we can track ships. An AIS transponder uses VHF frequencies, and broadcasts own-ship's position, name, call sign, along with detailed parameters like length, beam, draft, and tonnage. It also broadcasts details of the current navigation system: speed, course, rate of turn, destination, and ETA. More and more AIS base stations were deployed to receive this kind of real time ship information, and to send the information to VTS, where it will be displayed on the ECDIS [9] to facilitate the traffic monitoring.

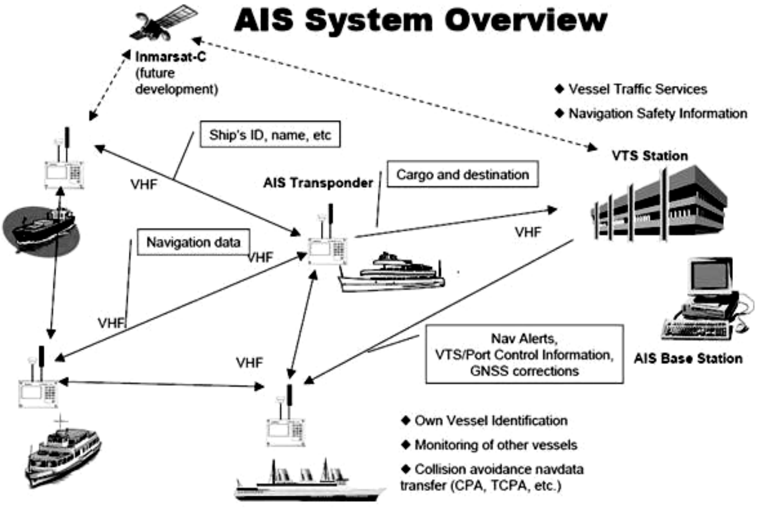

AIS System Overview
Inmarsat-C (future development)
Vessel Traffic Services
Navigation Safety Information
Ship's ID, name, etc
VTS Station
VHF
AIS Transponder
Cargo and destination
VHF
Navigation data
VHF
VHF
Nav Alerts, VTS/Port Control Information, GNSS corrections
AIS Base Station
VHF
Own Vessel Identification
Monitoring of other vessels
Collision avoidance navdata transfer (CPA, TCPA, etc.)

Fig. 2. System overview of AIS

The AIS data contains static data, dynamic data, and voyage-related information, security-related short messages and AIS ship reporting. In this paper, the static data and dynamic data are used. Static data includes the IMO identification number, IMO ship only code (MMSI), call sign and name of the vessel the length and width, vessel type and so on.

Dynamic data includes the ship's position (longitude, latitude), coordinated universal time UTC (dates generated by the GPS receiving equipment), true course, speed related to the ground, heading, status (for example, NUC, anchorage from the manual input), turnrate and so on.

2.2. Data Preprocessing

There are some data is the default, for example, speed is -1, longitude(lon) is 181, latitude(lat) is 91, heading is -1, length is 0, width is 0.

The treatment processes are as followings: if speed is -1, select the speed at previous time, if this speed is also -1, selects the speed at time before the previous time, and the selective processes continue until the speed is available. The treatment of other default is similar to speed.

In order to avoid conflict phenomenon when AIS data are issued, the time of data issued by different vessels is different in the same region. Therefore, so as to ensure the test data is simultaneous, the average time of test data is considered as the standard time, vessels' positions at this standard time are calculated by track reckoning.

3. CLUSTERING ALGORITHMS

There are two basic types of clustering algorithms [5] partitioning and hierarchical algorithms. *Partitioning algorithms* construct a partition of a database *D* of *n* objects into a set of *k* clusters. *k* is an input parameter for these algorithms, i.e some domain knowledge is required which unfortunately is not available for many applications.

The partitioning algorithm typically starts with an initial partition of *D* and then uses an iterative control strategy to optimize an objective function. Each cluster is represented by the gravity center of the cluster *(k-means algorithms*) or by one of the objects of the cluster located near its center (*k-medoid algorithms*). Partitioning algorithms use a two-step procedure. First, determine *k* representatives minimizing the objective function. Second, assign each object to the cluster with its representative "closest" to the considered object.

Ng & Han [6] explore partitioning algorithms for KDD in spatial databases. An algorithm called CLARANS (Clustering Large Applications based on

RANdomized Search) is introduced which is an improved k-medoid method. Compared to former k-medoid algorithms, CLARANS is more effective and more efficient.

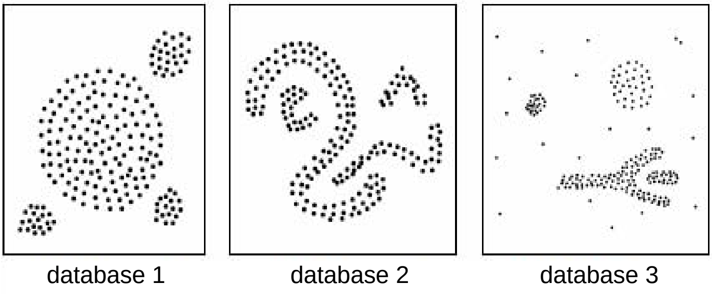
Hierarchical algorithms create a hierarchical decomposition of *D*. A hierarchical algorithms can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed. In such a hierarchy, each node of the tree represents a cluster of *D*. The dendrogram can either be created from the leaves up to the root (*agglomerative approach*) or from the root down to the leaves (*divisive approach*) by merging or dividing clusters at each step. In contrast to partitioning algorithms, hierarchical algorithms do not need *k* as an input.

4. DBSCAN ALGORITHM AND IMPROVEMENT

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density based clustering algorithm [4]. The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise.

4.1. A Density Based Notion of Clusters

When looking at the sample sets of points described in figure 3, we can easily and unambiguously detect clusters of points and noise points not belonging to any of those clusters.



database 1 database 2 database 3

Fig. 3. Sample databases

The density-based notion of clustering states that within each cluster, the density of the points is significantly higher than the density of points outside the cluster. Furthermore, the density within the areas of noise is lower than the density in any of the clusters.

Definition 1: The Eps-neighborhood of a point *p*, denoted by NEps (is defined by NEps(p) = $\{q \in D \mid \text{dist}(p, q) \leq \text{Eps}\}$. The distance function dist (p, q) determines the shape of the neighborhood. MinPts is the minimum number of points that must be contained in the neighborhood of that point in the cluster. There are two kinds of points in a cluster, points inside of the cluster (*core points*) and points on the border of the cluster (*border points*).

Definition 2: A point *p* is ***directly density-reachable*** from a point *q*. Eps, MinPts if
1) $p \in$ NEps(q) and
2) |NEps(q)| $\geq$ MinPts (core point condition).
Obviously, directly density-reachable is symmetric for pairs of core points. In general, it is not symmetric if one core point and one border point are involved. Figure 4 shows the asymmetric case.

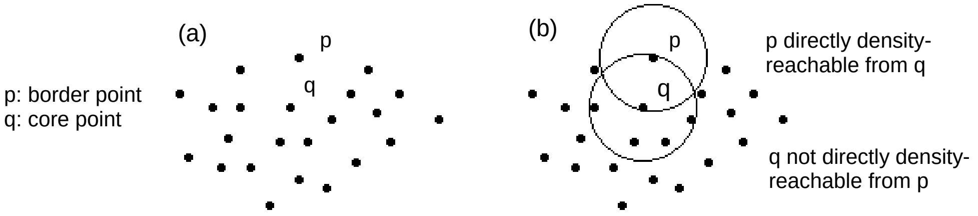

(a)
p
q
p: border point
q: core point
(b)
p
q
p directly density-reachable from q
q not directly density-reachable from p

Fig. 4. Core points and border points

Definition 3: A point *p* is ***density-reachable*** from a point *q*. Eps and MinPts if there is a chain of points $p_{1,}$ ..., p_n, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i. Density-reachability is a canonical extension of direct density-reachability. This relation is transitive, but it is not symmetric. Figure 3 shows the relations of some sample points and, in particular, the asymmetric case. Although not symmetric in general, it is obvious that density-reachability is symmetric for core points.

Definition 4: A point *p* is ***densityconnected*** to a point *q* wrt. Eps and MinPts if there is a point o such that both, *p* and *q* are density-reachable from o wrt. Eps and MinPts. Density-connectivity is a symmetric relation. For density reachable points, the relation of density-connectivity is also showed in figure 5.

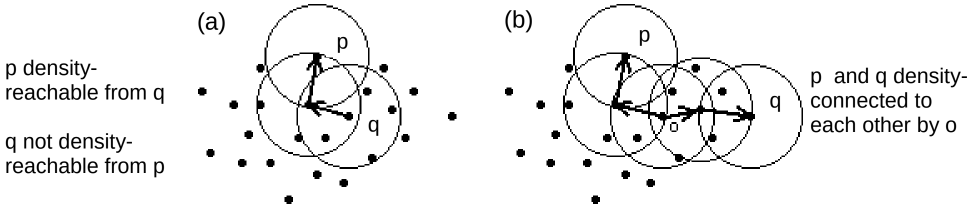

(a)
p
q
p density-reachable from q
q not density-reachable from p
(b)
p
o
q
p and q density-connected to each other by o

Fig. 5. Density-reachability and Density-Connectivity

Definition 5: (cluster) Let *D* be a database of points. A *cluster C* wrt. Eps and MinPts is a non-empty subset of *D* satisfying the following conditions:

1) ∀ *p, q*: if *p* ∈ *C* and *q* is density-reachable from p wrt. Eps and MinPts, then *q* ∈ *C*. (Maximality).
2) ∀ *p, q* ∈ *C*: *p* is density-connected to q wrt. EPS and MinPts. (Connectivity).

Definition 6: (noise) Let *C1,…, Ck* be the clusters of the database *D* wrt. parameters Epsi and MinPtsi, *i* = 1,…, *k*. Then we define the *noise* as the set of points in the database *D* not belonging to any cluster *Ci* , i.e. noise = {*p*∈*D* | ∀ *i*: *p* ∉ *Ci*}.

4.2. Improvement of DBSCAN Algorithm

In this paper, the shape of a ship will be simplified to a rectangular when clustering. And using simplified Fujii model instead of ε neighborhood in the DBSCAN algorithm, and in order to detect the water traffic congestion zones better, two different neighborhood models are defined based on simplified Fujii model. The three models are shown in Figure 6(a) (b) (c) are respectively.

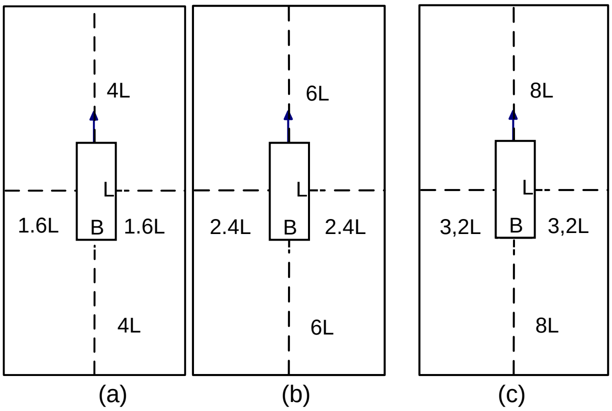


Fig. 6. (a)Neighborhood One (b) Two (c) Three

5. FUZZY EVALUATION MODEL OF MAIN TRAFFIC CONGESTION DEGREE BASED ON MAIN TRAFFIC FLOW VELOCITY

Fuzzy evaluation model for traffic congestion degree is built through combining traffic flow speed and fuzzy inference system, and the evaluation process is shown in figure 7.

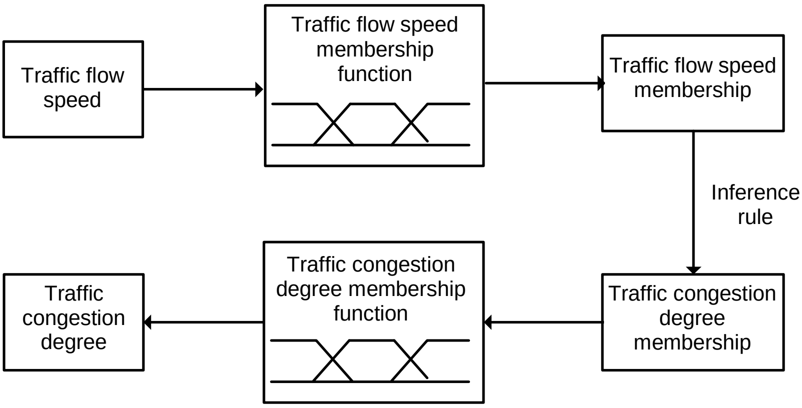

Traffic flow speed
Traffic flow speed membership function
Traffic flow speed membership
Inference rule
Traffic congestion degree membership
Traffic congestion degree membership function
Traffic congestion degree

Fig. 7. Fuzzy Evaluation Process for Traffic Congestion Degree

Traffic flow speed is divided into three grades: very slow, slow and fast, and the curve chart of traffic flow speed membership function is shown in figure 8. In addition, traffic congestion degree is also divided into three grades: uncongested, congested, very congested, and the curve chart of traffic congestion degree membership function is shown in figure 9.

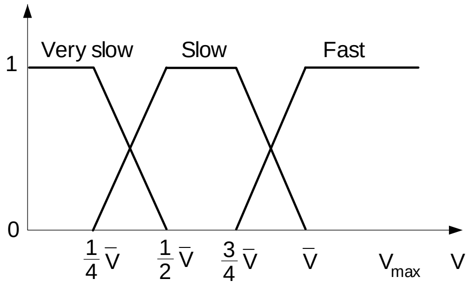


Fig. 8. Traffic Flow Speed Membership Function

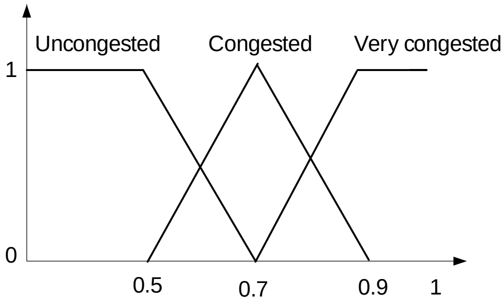


Fig. 9. Traffic Congestion degree Membership Function

5.1. Fuzzy inference rule of the evaluation

The fuzzy inference rule between traffic flow velocity and traffic congestion degree should be:

- if traffic flow velocity is “**very slow**”, then traffic congestion degree is “**very congested**”,
- if traffic flow velocity is “**slow**” then traffic congestion degree is “**congested**”,
- if traffic flow speed is “**fast**” then traffic congestion degree is “**uncongested**”.

6. EXPERIMENTAL EVALUATION

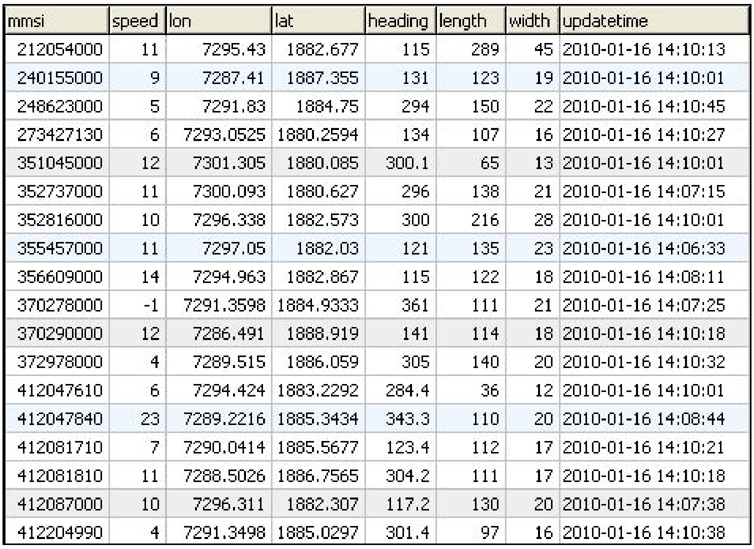
6.1. Dataset

We used database from Microsoft SQL Server database management software. It has following columns: mmsi, turntate, speed, longitude, latitude, course, heading, updatetime, width and length of all ships in the water channel.

6.2. Test Data and Preprocessing

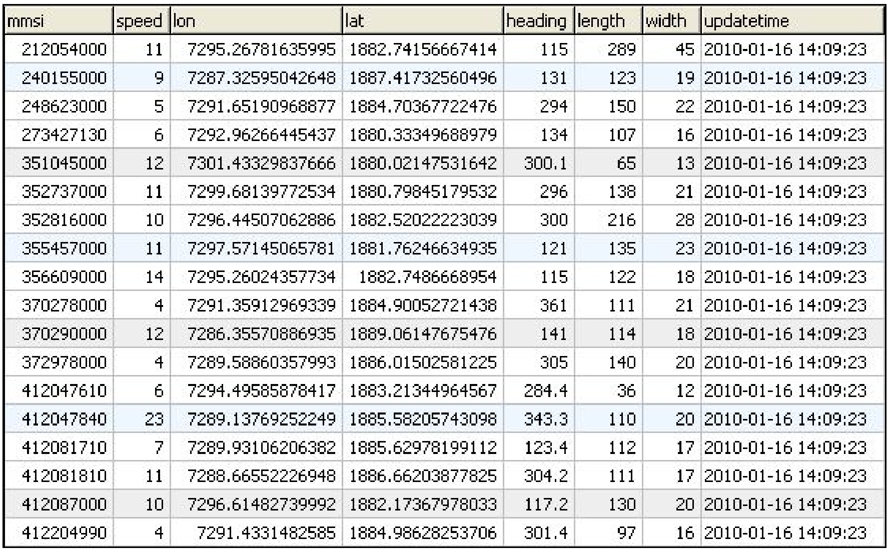
A large amount of traffic flow information, usually obtained by data collection equipment such as radar and AIS, are recorded and accumulated in the database of a Vessel Traffic Management System. However, the original data-set often includes noisy, missing and inconsistent data. Data preprocessing will improve the quality of the data and facilitate efficient data mining tasks.

Test data in this paper is the AIS data from the vessels nearby Wusongkou in Shanghai, as is shown in figure 10. And a part of test data after preprocessing is shown in figure 11.



mmsi	speed	lon	lat	heading	length	width	updatetime
212054000	11	7295.43	1882.677	115	289	45	2010-01-16 14:10:13
240155000	9	7287.41	1887.355	131	123	19	2010-01-16 14:10:01
248623000	5	7291.83	1884.75	294	150	22	2010-01-16 14:10:45
273427130	6	7293.0525	1880.2594	134	107	16	2010-01-16 14:10:27
351045000	12	7301.305	1880.085	300.1	65	13	2010-01-16 14:10:01
352737000	11	7300.093	1880.627	296	138	21	2010-01-16 14:07:15
352816000	10	7296.338	1882.573	300	216	28	2010-01-16 14:10:01
355457000	11	7297.05	1882.03	121	135	23	2010-01-16 14:06:33
356609000	14	7294.963	1882.867	115	122	18	2010-01-16 14:08:11
370278000	-1	7291.3598	1884.9333	361	111	21	2010-01-16 14:07:25
370290000	12	7286.491	1888.919	141	114	18	2010-01-16 14:10:18
372978000	4	7289.515	1886.059	305	140	20	2010-01-16 14:10:32
412047610	6	7294.424	1883.2292	284.4	36	12	2010-01-16 14:10:01
412047840	23	7289.2216	1885.3434	343.3	110	20	2010-01-16 14:08:44
412081710	7	7290.0414	1885.5677	123.4	112	17	2010-01-16 14:10:21
412081810	11	7288.5026	1886.7565	304.2	111	17	2010-01-16 14:10:18
412087000	10	7296.311	1882.307	117.2	130	20	2010-01-16 14:07:38
412204990	4	7291.3498	1885.0297	301.4	97	16	2010-01-16 14:10:38

Fig. 10. A Part of Test Data Before Preprocessing



mmsi	speed	lon	lat	heading	length	width	updatetime
212054000	11	7295.26781635995	1882.74156667414	115	289	45	2010-01-16 14:09:23
240155000	9	7287.32595042648	1887.41732560496	131	123	19	2010-01-16 14:09:23
248623000	5	7291.65190968877	1884.70367722476	294	150	22	2010-01-16 14:09:23
273427130	6	7292.96266445437	1880.33349688979	134	107	16	2010-01-16 14:09:23
351045000	12	7301.43329837666	1880.02147531642	300.1	65	13	2010-01-16 14:09:23
352737000	11	7299.68139772534	1880.79845179532	296	138	21	2010-01-16 14:09:23
352816000	10	7296.44507062886	1882.52022223039	300	216	28	2010-01-16 14:09:23
355457000	11	7297.57145065781	1881.76246634935	121	135	23	2010-01-16 14:09:23
356609000	14	7295.26024357734	1882.7486668954	115	122	18	2010-01-16 14:09:23
370278000	4	7291.35912969339	1884.90052721438	361	111	21	2010-01-16 14:09:23
370290000	12	7286.35570886935	1889.06147675476	141	114	18	2010-01-16 14:09:23
372978000	4	7289.58860357993	1886.01502581225	305	140	20	2010-01-16 14:09:23
412047610	6	7294.49585878417	1883.21344964567	284.4	36	12	2010-01-16 14:09:23
412047840	23	7289.13769252249	1885.58205743098	343.3	110	20	2010-01-16 14:09:23
412081710	7	7289.93106206382	1885.62978199112	123.4	112	17	2010-01-16 14:09:23
412081810	11	7288.66552226948	1886.66203877825	304.2	111	17	2010-01-16 14:09:23
412087000	10	7296.61482739992	1882.17367978033	117.2	130	20	2010-01-16 14:09:23
412204990	4	7291.4331482585	1884.98628253706	301.4	97	16	2010-01-16 14:09:23

Fig. 11. A Part of Test Data After Preprocessing

6.3. Clustering Results

The test data after preprocessing are analysed by the improved DBSCAN algorithm, and the MinPts is supposed to be 4.A part of results is shown in Figure three is about 3.78 knots, its congestion degree is 0.735, so it is congested. The traffic flow speed of class 2 in clustering results based on neighborhood three is about 5.01 knots, its congestion degree is 0.7, so it is congested.

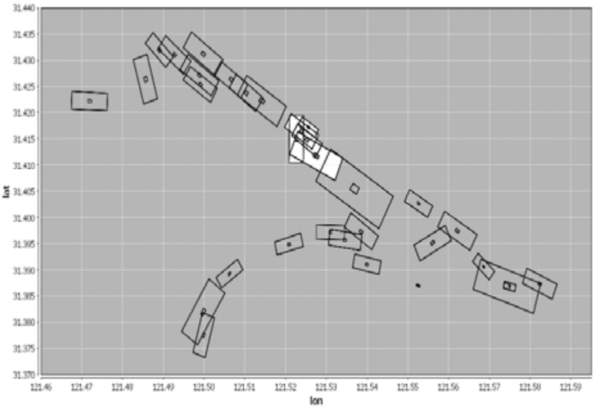


Fig.12. Congested Zones Visualization Based on Neighborhood One

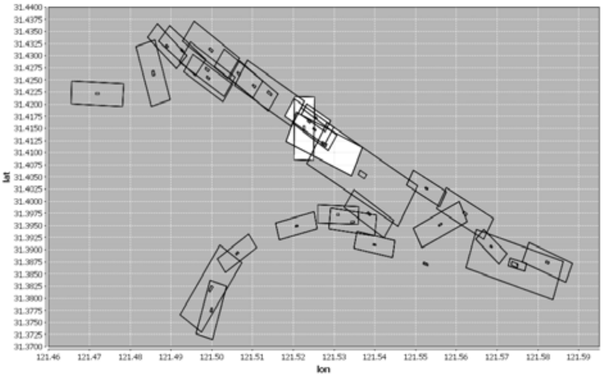


Fig. 13. Congested Zones Visualization Based on Neighborhood Two

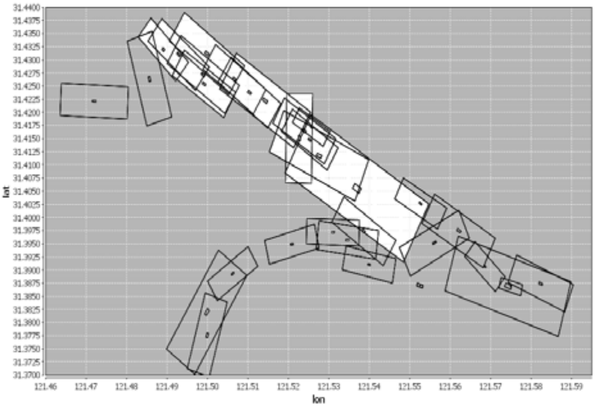


Fig. 14. Congested Zones Visualization Based on Neighborhood Three

CONCLUSIONS

The aim of this study is to identify congested zones by investigating the current marine traffic speed based on DBSCAN algorithm in the Wusongkou, Shanghai which is the chosen geographical area of research. A fuzzy inference model was built to determine the marine traffic congested zones under varying traffic flow speed. The neighborhood three models are proposed to identify the traffic congested zones better. Minpts in clustering algorithm and channel boundary information are not considered and the number of vessels inference on traffic congestion degree is not taken into account.

Results of the marine traffic simulation studies according to the current traffic situation show the traffic flow speed of class 1 is about 4.6 knots and the traffic congestion degree is 0.7 for neighborhood one and two. The traffic flow speed of class 1 is about 3.78 knots and the traffic congestion degree is 0.735 and class 2 is about 4.6 knots and the traffic congestion degree is 0.7 for neighborhood three.

This study is supposed on several future academic studies and collision avoidance at sea. We have plans to apply DBSCAN method to determine traffic congestion degree with ship domain for collision avoidance system.

ACKNOWLEDGEMENT

The research work in this paper is supported by Shanghai Leading Academic Discipline Project with grant number S30602.

REFERENCES

1. Davis G.B., Carely K.M., *Computational Analysis of Merchant Marine GPS Data,* CASOS Technical Report, Carnegie Mellon University, 2006(11).
2. Ester M., Kriegel H.P., Sander J., Xu X., *A density-based algorithm for discovering clusters in large spatial databases*, Knowledge Discovery and Data Mining (KDD'96), Portland, 1996(8).
3. Fujii Y., *Traffic Engineering at Sea,* Haiwen Hall, Tokyo 1981.
4. Han J., Kamber M., *Data Mining: Concepts and Techniques*, second edition, Morgan Kaufmann, 2006.
5. Kaufman L., Rousseeuw P.J., *Finding Groups in Data: An Introduction to Cluster Analysis,* Wiley, New York 1990.
6. Ng R.T., Han J., *Effcient and Effective Clustering Methods for Spatial Data Mining,* Proc. 20th Int. Conf. on Very Large Data Bases, Santiago, Chile 1994.
7. Qinyou H., Yong J., Shi Ch., Chen G., *Evaluation of Main Traffic Congestion Degree for Restricted Waters with AIS Reports,* 8th International Symposium on Marine Navigation and Safety of Sea Transportation, TransNav 2009.
8. Wang Shi-yuan, Xu Kai-yu., *Actuality, Prospect and Counterplan of AIS*, Marine Technology, 2001(10).

9. Weintrit A., *The Electronic Chart Display and Information System (ECDIS). An Operational Handbook.* A Balkema Book. CRC Press, Taylor & Francis Group, Boca Raton–London–New York–Leiden 2009.
10. Yong Jiajia, *First Order Real-time Safety Evaluation of Navigation Safety in Restricted Waters,* Shanghai Maritime University [dissertation], Shanghai 2009.
11. Yuan An-cun, Zhang Shu-fang, *International Standard Assembly of Ship Automatic Identification System,* Dalian Maritime University Press, Dalian 2005(4).

ANALIZA I IDENTYFIKACJA STREF NATĘŻENIA RUCHU MORSKIEGO W OKOLICACH PORTU WU SONG KU W SZANGHAJU

Streszczenie

Szanghaj, z jego naturalnym, kulturalnym i historycznym bogactwem, jest nie tylko jednym z najpiękniejszych miast chińskich, ale również jednym z najbardziej fascynujących miast na świecie. Jednakże wody wokół Szanghaju stanowią duże wyzwania dla nawigacji ze względu na skomplikowane geograficzne, geopolityczne i oceanograficzne struktury obszaru. Jednym z wyzwań jest ogromne natężenie ruchu morskiego, który krzyżuje się w tym miejscu. Wusongkou to miejsce zlokalizowane wzdłuż północnego końca rzeki Huanpu, która płynie z południowego zachodu na północny wschód i wpływa do rzeki Jangcy. W niniejszym opracowaniu podjęto próbę identyfikacji stref natężenia ruchu w okolicach Wusongkou opartą na algorytmie DBSCAN (Density-Based Spatial Clustering of Applications with Noise) i modelowaniu odległościowym z rozmytym kryterium dopasowania. W zastosowanych algorytmach DBSCAN uwzględniono domenę statku, a także algorytmy grupowania pozwalające częściowo uwzględnić rzeczywiste dynamiczne dane statków w celu wykrycia obszarów o największym natężeniu ruchu. Głównym celem pracy jest budowa modelu algorytmu predykcji, identyfikacji i prezentacji w czasie rzeczywistym obszarów o szczególnie dużym natężeniu ruchu morskiego dla rzeczywistych danych statystycznych o ruchu statków na akwenie podejściowym do Shanghaju oraz wybór modelu algorytmu optymalnego dla danego akwenu wodnego.