Clustering of Marine Vessel Trajectory Data for Routes Planning through Water Areas with Heavy Traffic

V M Grinyak\textsuperscript{1,2,3}, A V Shulenina\textsuperscript{2}, A S Devyatisilnyi\textsuperscript{3}

\textsuperscript{1}Department of Information Systems and Technologies, Vladivostok State University of Economics and Service, Gogolya, 41, Vladivostok, 690014, Russia
\textsuperscript{2}Department of Applied Mathematics, Mechanics, Control and Software, Far Eastern Federal University, Ajax village, 10, Russky island, Vladivostok, 690922, Russia
\textsuperscript{3}Department of Navigation and Control, Institute of Automation and Control Processes FEBRAS Radio, 5, Vladivostok, 690041, Russian Federation

E-mail: victor.grinyak@gmail.com

Abstract. The article is devoted to the problem of ensuring the safety of vessel traffic. One of the elements of the traffic organization in areas with heavy navigation is the system of establishing the routes of vessels. This system is a set of restrictions imposed by a certain traffic pattern and rules adopted in a particular water area. The paper considers the problem of planning a transition route for water areas with heavy marine traffic. The planning of the vessels transition route during the movement of the water area with established routes must be carried out taking into account the specified restrictions. A possible way to identify these restrictions is to isolate the movement patterns of a particular sea area from the retrospective information about its traffic. Model representations of such a problem can be formulated on the basis of the idea of clustering the parameters of traffic. The route planning problem model is based on finding the shortest path on a weighted graph. Several ways of constructing such a graph are proposed: a regular grid of vertices and edges; a layered or random grid of vertices and edges; vertices and edges based on retrospective data. The weight of the edges is proposed to be set as a function of the “desirability” of a particular course of the vessel for each point of the water area, taking into account the identified movement patterns. The paper discusses possible clustering methods and makes a choice in favor of subtractive clustering.

1. Introduction

The use of new methods for solving classical problems of navigation and the study of new problems has become possible due to the development of methods and tools for collecting, storing and processing large amounts of data, services of the Automatic Identification System (AIS), e-navigation technologies. One of these new problems is the assessment of characteristic patterns of sea traffic, the methods of solving which are based on the ideas of clustering.

Information about traffic patterns can be used when dispatching traffic in a given area, for example. Another prospective area is application of information about specific parameters is planning the transition route through waters with heavy traffic. The essence of the problem lies in the fact that the features of movement on a certain sea section, where the intended route of the vessel is located along, render a significant influence on the choice of the route from the set of possible ones. For example, the system for establishing the routes of movement of vessels, which is a set of restrictions due to the
scheme of movement of vessels ("traffic rules") defined for a particular water area, is often introduced for water areas with heavy traffic. The described traffic pattern based on many years of collective navigation experience in a particular area can be adopted both explicitly and implicitly (informally). When a vessel moves through such areas, the planning of the transit route must be carried out taking into account the restrictions imposed.

The classical problem of planning a transit route is an optimization problem, the purpose of which is to navigate along the shortest possible path, in the shortest possible time, with the lowest fuel consumption, etc. [1], [2]. For the considered variant of the route planning problem, taking into account the constraints, the mathematical model can be based on the search for the shortest path on the weighted graph of possible ship routes. This approach is traditional and well proven in this class of problems [3]. The weight of the graph edges is determined by the "desirability" of a particular vessel's course for each point of the water area, taking into account the identified movement patterns. As a result, the most typical, and therefore safe route for the selected water area is selected. The considered problem of planning ship traffic routes based on information about the characteristic traffic is topical. This is connected with the prospects for the development of unmanned navigation.

2. Basic model representations

Let’s introduce the \( x, y \) coordinate system, where the \( x \)-axis is the geographic longitude, and the \( y \)-axis is the geographic latitude of the vessel. We determine the desirability function of the parameters of the vessel’s motion \( u(x, y, k, v) \) where \( k \) is the course, \( v \) is the speed, \( x, y \) are the coordinates of the vessel. The desirability of the vessel's movement along the chosen route \( q \) (some curve) can be expressed as a curvilinear integral of the first kind

\[
\int_{q} u(x, y, k, v) dq.
\]

The solution to the optimization problem of choosing a route \( q \) ensures the minimum of the functional \( U \):

\[
q_* = \arg\min_{q} U(q).
\]

A direct solution to the latter problem is in principle possible. However, in practice, it is rarely used because of the high computational complexity and (as a rule) the discreteness of the grid of arguments of the function \( u(x, y, k, v) \). A popular heuristic simplification of the problem is its model interpretation by finding the shortest path on the weighted graph of possible routes of the vessel [3].

Consider the approaches to building such a graph.

A regular grid of vertices. Dividing the water area into square sections.

Let’s divide the water area into many square sections. Let’s select the points of the centers of the squares in the areas in which the movement of the vessel is permissible. Let us assume that these points of the square centers form the set of vertices of the graph of possible routes of the vessel. One of them corresponds to the starting point of the vessel’s movement, the other to the final point of the route. The set of edges of such a graph can be defined by connecting vertices according to the "each with each" principle, excluding those edges that pass through sections that are unacceptable for movement. An additional condition for the maximum length of the graph edges should be introduced to reduce their number. The absence of restrictions on the amount of change in the ship’s course at the points of the vertices is a disadvantage of such a graph. This can lead to inconvenient or even unrealizable routes.

A regular grid of vertices. Layered division of the water area.

Let’s assume that the possible routes of the vessel are inside the rectangle. The rectangle is constructed in such a way that the start of the route is in the middle of one side of the rectangle, and the end of the route is in the middle of the opposite side. The rectangle is filled with the vertices of the
graph so that a regular rectangular grid is formed. Vertex layers are the vertices of the graph that lie on segments perpendicular to the segment with vertices at the start and end points of the route. When the ship is moving, only transitions between the nearest vertex layers are possible. Edges that pass through sections that are not allowed to move are excluded. It is also necessary to limit the maximum length of the edge. This representation of the graph takes into account the navigational specifics of the movement and implicitly limits the amount of course change.

A random grid of vertices.

Let's take as a basis the division of the water area described above into square areas. Let the probability \( P(x,y) \) be given that the point of the center of the squares with coordinates \( x, y \) is the vertex of the graph of possible vessel routes. This probability may be the same for the entire water area, or it may be different for different areas. For example, it may increase in the places where ship traffic intersects and in the areas of characteristic course changes. In the degenerate case \( P(x,y)=1 \), the set of vertices will coincide with the regular grid. The set of edges of the graph is given in the same way as for partitioning into square sections. The problem of a random vertex grid is the difficulty of formally justifying the function \( P(x,y) \). It seems that it can only be set in a heuristic way. In addition, different implementations of the graph of possible routes, when randomly generated, can lead to significantly different results for finding the best route. On the other hand, the subsequent selection of the "final" best route by some criterion is possible after several implementations.

A grid of vertices based on retrospective data.

Let there be data about the movement of vessels in the selected water area for a certain period of time in the form of a set of tuples of the values of longitude, latitude, speed and course of the vessel. Let's define the set of vertices of the graph of possible routes based on the data of this set. In this case, there are two main options. In the first variant, the set of vertices of the graph is formed on the basis of the complete set of data on the coordinates of ships. In the second option, only a part of the set is selected. For example, as in the described approach with a random grid of vertices. The set of edges of the graph can be defined in the same way as when dividing into square sections, or it can be formed on the basis of motion data, if subsets of tuples belonging to the same trajectory are known. In the latter case, it should be supplemented with "not implemented" edge options to ensure the graph connectivity. The advantage of this approach is the formation of possible ship routes that are specific to a particular water area. The disadvantage is the difficulty of formally justifying the time interval for which the traffic data is taken. It seems that it should be chosen in a heuristic way.

Consider the approaches to setting the weight of the edges of a graph of possible routes. Assume that the weight of the edge is equal to the length of the arc of the great circle connecting the incident edges of the vertex in the absence of data on the movement of ships. The weight coefficient of each edge \( a \in [0, 1] \) is entered if there is data on the movement of ships. This coefficient is set in one of the following ways.

Accounting for the number of vessels with a close course and speed.

Let there be retrospective data on the movement of ships (a set of tuples of longitude, latitude, speed, and course values) in some neighborhood of the selected edge. Let us calculate the number of \( m \) “close” (that is, lying in a certain range of courses and speeds) vectors of the ships' speed in the vicinity of the selected edge, setting the course corresponding to the edge and the assumed speed. The weight coefficient of the edge is assumed to be \( a=1/m \). Thus, the more vessels previously moved along the corresponding edge of the trajectory, the easier it will be (and, accordingly, preferable when planning the route). It is also possible to calculate \( m \) only on the course data, without taking into account the estimated speed of the vessel. The disadvantage of this approach is that all ships in the vicinity of the edge are counted, without taking into account their routes. This can lead to the fact that the planned route will go along the most "popular" edges, even if they are not preferred in a particular case.

Accounting the characteristic values of the course and speed.

In order to exclude the effect of "popularity", it is necessary to determine the "specificity" of the values of courses and speeds without explicitly taking into account the number of vessels
implementing them. In this case, the idea of clustering is productive. The set of characteristic values of the velocity vector of ships located in the vicinity of the selected edge-clusters of velocity vector values is determined from retrospective data. We determine whether the specified velocity vector belongs to one of the clusters by setting the corresponding edge course and estimated speed. In this case, the weight coefficient of the edge $a$ is set to be small (for example, $a=0.1$). Thus, when planning a route, preference will be given to those edges that correspond to the characteristic movement. It is also possible to determine clusters only based on the course data, without taking into account the estimated speed of the vessel.

**Edges based on retrospective data.**

This method of determining edge weights can be used if the vertex grid is formed based on retrospective data (see above). For edges formed on the basis of motion data, the weight coefficient of the edge is set to a small value (for example, $a=0.1$). For those edges that complement the original set, making the graph connected, the weight factor is assumed to be large (for example, $a=1$). Thus, routes that have already been implemented become preferable.

The search for the shortest path with a small number of vertices and edges of the graph of possible ship routes can be carried out by well-known deterministic algorithms (Dijkstra, Bellman-Ford, etc.). The complexity of the most efficient deterministic algorithms is proportional to the number of edges and the number of vertices (or their logarithm). The number of vertices and edges of the graph may be too large for the productive work of deterministic algorithms if large samples of retrospective data are used for its formation. In this case, heuristic algorithms are perspective: ant and genetic.

The set of properties of the considered approaches to the construction of the graph of possible vessel routes allows us to assume that the layered partition of the water area by a regular grid of vertices is the most suitable for the considered problem of planning a route through water areas with heavy traffic. It has also proven itself well in other tasks [1]. The weight of the edges is set taking into account the characteristic values of the course and speed determined by clustering the movement parameters.

A metric is introduced for clustering. A metric is a function that sets the degree of proximity between objects. In the problem under consideration, this metric of the distance between objects 1 and 2 can be entered as follows:

$$D_{12}^2 = w_{\text{lon}}(x_1 - x_2)^2 + w_{\text{lat}}(y_1 - y_2)^2 + w_{\text{speed}}(v_1 - v_2)^2 + w_{\text{course}}(k_1 - k)^2.$$  

Here $w_{\text{lon}}, w_{\text{lat}}, w_{\text{speed}}, w_{\text{course}}$ are weight coefficients that are set based on the data on the characteristic sizes of clusters for each of the dimensions. The course difference function takes into account the periodicity of the angle data. Determining weights by the coordinates $w_{\text{lon}}, w_{\text{lat}}$ is a non-trivial task. This is due to the fact that in the water area there can be both sections of maneuverable movement of several hundred meters in size, and zones of long-term uniform movement of several tens of kilometers. Therefore, for some applications (including the problem under consideration), you can resort to decomposition. It is possible to perform clustering on separate selected sections of the water area only by speed and/or course.

Mountain and subtractive clustering, which do not require specifying the number of clusters, seems to be suitable for estimating the characteristic values of the course and speed. The mountain clustering algorithm is as follows. Let there be a set of $M$ objects and a distance matrix $D_{ij}$ that specifies the degree of proximity between objects with indices $i$ and $j$. Let us assume that the objects themselves are possible centers of clusters. For each object, the value of its potential is calculated as follows:

$$p_i = \sum_{j=1}^{M} \exp(-aD_{ij}),$$
where $\alpha$ is a number characterizing the scale of distances $D_{ij}$, $\exp(\ )$ is an exponent operator. At the first step of the algorithm, an object with the index $max_1$, which has the maximum potential, is selected. This object will be the center of the first cluster. In the second step, the values of the potentials of the objects are recalculated according to the formula:

$$p_{i}^{(2)} = p_i - p_{max_1} \exp(-\beta D_{ij}),$$

where $\beta$ is the number that determines the size of the clusters. The point with the index $max_2$ with the maximum value of the potential $p_{i}^{(2)}$ will be the center of the second cluster. The centers of all the following clusters are located in the same way. The procedure continues as long as the potential of the next cluster exceeds a certain threshold, or as long as there is a difference between the potentials of neighboring levels, for example.

Subtractive clustering is similar to mountain clustering. After finding the center of the first cluster, those that belong to it are excluded from the set of objects. The procedure is iteratively repeated until the value of the potential of the next cluster exceeds the specified threshold.

3. Results of calculations based on traffic data

The study was carried out for a certain number of water areas using real data on the movement of vessels collected from the marinetrack.com resource using a specially developed software system. Here is an example of a successful solution to the problem for the Tsugaru Strait and Tokyo Bay. In the first case, to find the route of ships through the water area, traffic data was taken for one week (in total, about 1.5 million records of the "longitude, latitude, speed, course" type). In the second case – within 3 days (about 2 million records in total). The graph of possible routes of the vessel was formed on the basis of layered division of the water area by a regular grid of vertices. The partition parameters were selected in such a way that the length of the graph edges did not exceed 3 kilometres. The weight of the edges was set taking into account the characteristic values of the course on the segment $[0, 360^\circ)$, determined by clustering the movement data in the vicinity of each edge (Fig. 1). If the movement along the edge did not correspond to the course typical for a section of the water area, its weight was assumed to be equal to the length of the arc of a large circle. Otherwise, it was multiplied by the weight factor $a=0.1$.

![Figure 1. Graph of possible ship routes. The weight of the edges is determined based on the movement data in the vicinity of the edge (the areas highlighted in gray).](image)

The standard method of subtractive clustering was used. The following parameters of subtractive clustering were set. The parameter $\alpha$ corresponds to the cluster radius of $16^\circ$. Objects belong to a cluster if they lie closer than $20^\circ$ from its center ($1.25 \alpha$). The procedure for searching for cluster centers continues until the potential of the next cluster exceeds 10% of the potential of the first cluster. Such values of the method parameters correctly identified the clusters of the reference data sample.

Figure 2 shows the results of planning the routes Pacific Ocean – Sea of Japan (east – west) and Hakodate Port – Mutsu Bay (north-south) and back. Figure 3 shows the planning of the route Pacific Ocean – Yokohama and back. It can be seen that the routes found fully correspond to the direction of
movement of real ship traffic, take the "right" side of the movement (shown by the arrows, right-hand traffic is accepted in navigation).

4. Discussion
In connection with the development of new information technologies and the concept of e-navigation, there has been an evolution of traditional models of navigation tasks in the direction of describing the collective movement of vessels in the water area as a whole. We can note the work [4]. In this work, the number of dangerous approaches of vessels of various classes in certain areas of the water area is estimated on the basis of the classical model of the ship's domain and the most difficult areas for movement are identified. Studies [5], [6], [7] are devoted to the same problem.

![Figure 2. Results of route planning in the Tsugaru Strait.](image1)

![Figure 3. Results of route planning in Tokyo Bay.](image2)

The paper [4] proposes an approach for assessing the traffic intensity in the water areas as a function of the density of ships. The study [5] proposes a method for estimating traffic intensity that takes into account the geometric dimensions of the vessels. It is noted that some water areas have
limited capacity for large vessels. There is an example in which for the safe passage of large vessels, it is necessary to reduce their number. The article [6] suggests an approach to choosing the speed of a ship based on retrospective data on the traffic of a particular water area (the article analyzes the movement of ships near Shanghai). The speed is selected depending on the traffic density. Thus, the collective experience of navigators is implicitly taken into account.

Approaches related to the cluster analysis of data on the vessels movement in the water area are also becoming increasingly developed. They explore options for clustering objects, their attributes, metrics, and clustering methods. Thus, the paper [7] considers the problem of estimating typical ship routes in the water area according to AIS data. The method is based on the division of the water area into small sections, the assessment of the traffic density in them, the preferred transitions of ships between them, i.e. clustering occurs implicitly. As a result, polylines are formed, which represent the routes of ships between the specified start and end points, taking into account the previous traffic, including the accepted scheme of ship traffic. A feature (and, apparently, a significant disadvantage) of the method is the need for subsequent smoothing of the obtained routes.

In work [8], clustering is also used to estimate typical water area routes. The objects of clustering are polylines formed from retrospective AIS data. The metric of the distance between the polygonal lines is introduced. The clustering algorithm is based on finding areas of connectivity. A method for selecting the parameters of the algorithm is proposed. As a result, the main routes of ships in the area of intensive navigation are identified (an example for one of the regions of the South China Sea is given in the work). This makes it possible to identify abnormally moving vessels whose route is not typical for the given water area. The method does not solve the route planning problem. A similar problem is considered in the article [9]. The objects of clustering are data on the coordinates, speeds, and courses of ships. The work [10] also considers the problem of identifying abnormally moving vessels, the signs are their coordinates, courses and speeds. The water area is divided by a rectangular grid, the routes of ships are represented by the rules of passage between the grid cells. Clustering is based on the construction of histograms of courses and speeds for each cell according to AIS data. Abnormal values of speeds and course are identified by these histograms. Although the approach proposed in this paper can be used as the basis for the task of planning the ship's route, it is not explicitly considered. The disadvantages of the method are the need to pre-process the AIS data to build a route graph, the high computational complexity of the corresponding algorithms, and the need for a large sample of data to build histograms.

This work is also devoted to the application of cluster analysis of data on the vessels movement in the water area for solving the classic problem of navigation - planning the transition route. This approach provides a new quality of the route-its compliance with the established traffic parameters developed by collective professional experience. Postulating that the retrospective data on movement in the water area are the result of positive operational practices, it can be assumed that the route planned on their basis will be the safest possible in specific navigation conditions.

The mentioned method [10], based on the determination of traffic rules by statistical methods, is the closest to the approach proposed in this paper, where it is proposed to use clustering methods. This makes it possible to reliably identify the motion parameters and does not require large amounts of initial data. So, in the one shown in Fig. 2, 3 examples, the characteristic amount of data on the course of ships in the "popular" areas of the water area was equal to 20-50, at the most saturated, as a rule, did not exceed 200. This made it possible to build a stable picture of the characteristic parameters of the movement. The variant of the graph of possible routes with layered division of the water area (Fig. 1) should be considered successful. On its basis, routes are formed that do not allow excessive changes in the course. The characteristic number of vertices and edges of the graph is small, even for extended routes. This allows us to limit ourselves to deterministic methods of finding the shortest path.

5. Conclusion

The paper deals with the problem of planning the route of a vessel's transition through water areas with heavy traffic. An approach based on the use of the idea underlying big data technology is
proposed. This idea is that the movement of the vessel should correspond to the characteristic kinematic parameters of movement, determined from retrospective traffic information. This makes it possible to take into account the collective experience of navigation in a particular water area. Model representations of the problem include a graph of possible ship routes and a function of the "desirability" of traffic parameters. The paper considers several possible variants of the model. The choice in favor of a layered division of the water area by a regular grid of vertices is justified. The weight of the graph edges is determined based on the results of clustering of retrospective motion data.

The method of constructing a graph of possible ship routes allows you to reduce the dimension of the data during clustering. The water area is divided into small sections. Clustering of traffic data is performed for each of the sections separately. The signs of the objects are the courses and speeds of movement (together or separately). It is indicated that it is preferable to use clustering methods that do not require pre-setting the number of clusters. For example, mountain and subtractive clustering methods can be used.

The services of the Automatic Identification System can serve as a source of data on water area traffic. The paper shows the possibility of using AIS data available on specialized Internet resources. Despite the "sparseness" of these data samples, they adequately represent the aggregate traffic features.

The paper provides examples of planning the transition route through the Tsugaru Strait and Tokyo Bay. It can be seen that the found routes correspond quite well to the characteristic movement of vessels in the water area. This fact confirms the prospects of the proposed approach for practice.

6. References
[1] Veremey E I and Sotnikova M V 2016 International Journal of Open Information Technologies 4(3) 55-61
[2] Akmaykin D A, Bukin O A, Grinyak V M and Moskalenko M A 2018 Marine intellectual technologies 4–5(42) 148-152
[3] Akmaykin D A and Grnyak V M 2019 IOP Conference Series: Earth and Environmental Science 272 032103
[4] Liu Z, Wu Zh and Zheng Zh 2020 Journal of Navigation 73(3) 628-645
[5] Weng J, Liao Sh and Yang D 2020 Journal of Navigation 73(1) 75-91
[6] Wang L, Li Y, Wan Zh, Yang Z, Wang T, Guan K and Fu L 2020 Ocean Engineering 204
[7] Naus K 2019 Journal of Navigation 73(3) 726-745
[8] Zhao L and Shi G 2019 Journal of Navigation 72(4) 894–916
[9] Zhen R, Jin Y, Hu Q, Shao Zh and Nikitakos N 2017 Journal of Navigation 70(3) 648-670
[10] Tang H, Wei L, Yin Y, Shen H and Qi Y 2019 Journal of Navigation 73(5) 1014–1035

Acknowledgments
This work was supported by the Russian Foundation for Basic Research, project no. № 20-38-90018.