

Anomaly Detection in White Shipping

By David Nevell

QinetiQ, Malvern Technology Park, St. Andrews Road, Malvern, UK, WR14 3PS

Abstract

Anomalous behaviour in white shipping includes that relating to its kinematic characteristics, such as course and speed. However, monitoring the possible threat posed by the worldwide movement of tens of thousands of ships between thousands of ports renders a manual process impractical. There is therefore a need to develop methods of automatically detecting possible anomalies from whatever movement and other relevant data is available. These can then be prioritised for further investigation. The mathematical challenge is one of finding a sufficiently rigorous and robust approach to anomaly detection that is computationally scalable to a global context. Although there are established shipping lanes, the ocean is a continuum and the effect, for example, of temporary weather systems may have a significant effect on routes taken, as well as a wide range of other factors including the class of ship, its cargo and its crew. This paper concentrates on network design and the application of Bayesian methods to detect kinematic anomalies within the context of that network such as changes of destination, and inconsistent and unexpected routings, for example, not appearing to be heading for a port.

1. Introduction

The methodologies described in this paper are the result of ongoing work under the Multi-Intelligence Techniques (MIT) project sponsored by DTIC. The problem of identifying anomalous white shipping behaviour currently has high relevance to MOD.

The principle of the approach is to develop a pre-defined network to model the global movement of ships, use it combined with appropriate cost functions to predict expected routes between parts of the network, and then compare actual ship movements with expected routes to assess whether the behaviour is anomalous. Although the approach can be substantiated with reference to individual ship past history, it does not depend upon this information being available. The methodology described below assumes that each successive observation of a ship includes its lat/long position and its heading.

Networks, consisting of discrete nodes and branches, have proven to be particularly suitable frameworks for modelling the movement of traffic in the land domain (e.g. Ref 1). Vehicles are constrained to travel along roads, junctions of roads represent driver decision points, and once a vehicle exits a junction, unless it stops or performs a U-turn, it is known at which junction it will arrive next. Under these circumstances the junctions can be represented by nodes and the roads can be represented by branches. In contrast, the ocean domain is a continuum. Nodes and branches are not limited in the same way as they are on land and in general, ocean journeys are constrained only by land masses. In these circumstances it is usually advantageous for ships to travel by the shortest available route, namely along segments of Great Circles. However, this is affected by a large number of other factors such as weather, sea conditions, exclusion zones and depth of water, not to mention more unpredictable behaviour due to a ship's captain's preferences.

One way of "discretising" the ocean space could be to define a grid of nodes spread across the surface of the world's oceans with a density sufficient for a specified accuracy. This would

have the advantage of ensuring that each vessel would always be close to a node but the considerable disadvantage of generating a huge number of nodes; 100km² grid squares would conservatively imply around 3 million nodes and 4 million branches. If however, the principle of a road network is adopted, whereby nodes are placed only at route decision points, then a much sparser network can be generated. The assumption of Great Circle route choice implies that all decision points will be close to landmasses, for example prominent convex land features that ships have to sail round. This type of network will, in comparison to land networks, be branch rich and node light. There are computational advantages in adopting such an approach as the speed of optimal route algorithms (e.g. Dijkstra's) are dominated by the number of nodes rather than the number of branches. Since the ultimate aim is to produce a global network, calculation time, both for network preparation and for analysis is an important issue. A node-sparse network might only need 20000 nodes and have journeys that involve 1-10 branches rather than hundreds.

2. Network Design

The approach that was adopted for network design is summarised in Fig 1. Each landmass in the world, whatever size, can be represented by a closed polygon. Lat/long coastline coordinate data, for a wide range of resolutions, is available at Ref 2. This was used to produce three different types of network node for each landmass.

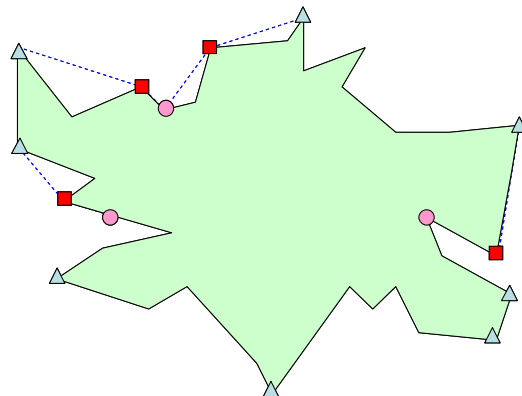


FIGURE 1. Different types of network node.

Firstly, all ports are added as network nodes (marked as circles). These are the only coastal network nodes that are considered to be valid start and finish points of a journey. Secondly, all additional coastal points that form part of the convex hull of the polygon are included as network nodes (triangles). Finally any other additional convex coastal nodes that may be required to delimit routes from ports to open water in a port or starboard direction are added (squares). In addition, there may also be a requirement for some offshore nodes, either potential destinations (e.g. oil rigs) or defining the edge of restricted routes (e.g. through the English Channel).

Great Circle Branches are then generated between pairs of nodes that are not impeded by any intervening landmass. In addition, branches that are expected to be of no practical use in route planning are also excluded (Fig 2). Whereas all sea-bound branches from Port A are considered admissible, the only ones that would be included from Node B (not a port) are those lie within one of the two inscribed angles shown. Other branches approach B at steeper angles that could not possibly lie on any optimal route and are therefore excluded.

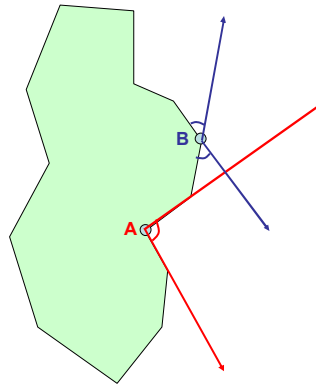


FIGURE 2. Admissible branch directions from coastal nodes.

Once the nodes and branches have been established, Dijkstra’s algorithm is used to pre-calculate optimal routes and costs between ports in the network. The baseline cost function for each branch is assumed to be its length. If necessary, the cost function can be augmented by canal tolls or other branch specific features. Pre-calculation of optimal routes allows subsequent analysis of ships’ routes to be speeded up considerably.

3. Route Analysis

A ship’s journey begins at a port and ends at a port and is assumed to follow a route which passes through or close to a sequence of intermediate network nodes. A complete journey J can be described by its constituent branches between those nodes such that

$$J = \{b_1, \dots, b_n\}$$

where the b_i represent branches. In practice, it is necessary to analyse incomplete journeys in order to assess whether the ship’s behaviour has been anomalous or not. An incomplete journey is made up of two parts, the macro part and the micro part. The macro part consists of all branches traversed before the previous network node. The micro part consists of the route since the previous network node. It is useful to divide the route in this way as the macro route can be described with reference to the pre-defined network, whereas the micro route, having no local nodes with which to identify, has to use a different and more precise frame of reference.

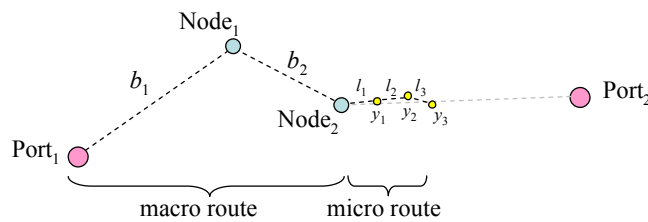


FIGURE 3. The macro and micro elements of an incomplete journey.

In the case shown (Fig 3), the intended journey is assumed to be from Port₁ to Port₂. So far the ship has traversed branches b_1 and b_2 (the macro route), and since the previous node Node₂ has been observed at points $y_1 \dots y_3$, implying legs $l_1 \dots l_3$ (the micro route). The underlying requirement is to calculate, for every possible destination D_i , the probability that the true destination is D_i , conditional on having observed the route so far. In the particular case above this can be expressed as

$$P_i = P(\text{Destination} = D_i | \text{Route so far} = \{b_1, b_2, l_1, l_2, l_3\})$$

Or more generally

$$P_i = P(D_i | R_M, R_m)$$

where R_M represents the macro route and R_m represents the micro route. Using Bayes Theorem and making the prior assumption that all ports are equally likely destinations, it can be shown that

$$P_i \propto P(R_m | D_i) P(R_M | D_i) \quad (1)$$

The macro element of expression (1) can be successively expanded to give

$$P(R_M | D_i) = P(\{b_1, b_2, \dots, b_q\} | D_i) = P(b_1 | D_i) P(b_2 | D_i) \dots P(b_q | D_i)$$

where $P(b_j | D_i)$ is the probability of next traversing branch b_j , conditional on the true destination being D_i . Consistent with land network terminology, these are referred to as exit probabilities. Whilst there are effectively an infinite number of directions that a ship may depart a node from, the exit probabilities in the ocean domain are restricted to the branches that have been pre-defined in the network.

The optimal route from any node conditional on the destination being D_i , has already been pre-calculated and if there was no uncertainty about whether ships followed the optimal route, the exit probabilities would all be in the set $\{0,1\}$ with only the branch on the optimal route being assigned a probability of 1. It is however, based on typical ship behaviour, more realistic to give slightly suboptimal routes non-zero probabilities. This is done in the following way.

Assume that to get to D_i , the optimal node to visit after n_{k-1} is n_{opt} and that the total cost of the journey from n_{k-1} to D_i is c_{opt} . Now, for any other possible next node n_{alt} , the cost of the trip to D_i via n_{alt} will be

$$c_{alt} = \text{cost}(n_{k-1} \text{ to } n_{alt}) + \text{cost}(n_{alt} \text{ to } D_i)$$

The likelihood of taking the route to D_i via n_{alt} can be expressed as:

$$L_{alt} = \exp(-\alpha(c_{alt} - c_{opt}))$$

where α is a constant that can be set to reflect the level of aversion to additional cost. Calculation of these likelihoods for all possible alternative next nodes allows the exit probabilities for all possible next nodes to be derived through normalisation, and therefore the macro element of expression (1) can be evaluated.

Calculation of the micro route probability, $P(R_m | D_i)$, utilises the heading of the ship at each observation. Expansion of this term, allowing for the fact that the micro route may be following any one of a number of different branches from the previous node, gives:

$$P(R_m | D_i) = \sum_{b_j \in \{B_{k-1}\}} P(h_1 | b_j) P(h_2 | b_j) \dots P(h_t | b_j) P(b_j | D_i)$$

B_{k-1} is the set of all branches emanating from the previous node n_{k-1} and h_1, \dots, h_t are the headings observed at each observation since the last network node.

The last term in the summation expression is an exit probability. Calculation of the $P(h_i | b_j)$, which will be referred to as heading probabilities, are based on the assumption that ships travelling on a heading h_i will typically exhibit deviation ϕ that follows a Gaussian distribution with mean 0 and variance σ_h^2 , i.e.

$$\phi \sim N(\theta : 0, \sigma_h^2)$$

The probability $P(h_i|b_j)$ of observing a heading h_i conditional on following branch b_j is therefore proportional to

$$\frac{1}{\sigma_h \sqrt{2\pi}} \exp\left(-\frac{(h_i - r_i)^2}{2\sigma^2}\right)$$

where r_i is the required heading to follow branch b_i . This then allows the micro route element of expression (1) to be calculated and hence the overall likelihood of the destination being D_i , based on the entire route observed so far. Similar likelihoods are calculated for every destination $D_i \in D$, the set of all possible destinations. These are then normalised to sum to unity, giving the required probability for each destination. D will contain every port. In addition, it may contain other specifically defined destinations (e.g. a fishing field or a rendezvous point).

4. Kinematic Anomaly Statistics

After every new observation of a ship's position, conditional destination probabilities are updated for every destination in the network. These probabilities can be used to assess the following two hypotheses.

H_0 : The ship is travelling to its stated destination

H_1 : The ship is travelling somewhere else

Implicit in the null hypothesis H_0 is the assumption that if a ship is travelling to its stated destination, then it will do so by an acceptable route, i.e. one that does not incur unreasonable cost.

The anomaly statistic of interest is $P(H_0|\text{route}_t)$ where route_t is the complete route covered by the ship up to time t . Using Bayes rule this can be written as

$$A_t := P(H_0 | \text{route}_t) = \frac{P(\text{route}_t | H_0)P(H_0)}{P(\text{route}_t | H_0)P(H_0) + P(\text{route}_t | H_1)P(H_1)}$$

Prior values for $P(H_0)$ and $P(H_1)$ are required. If it is observed, for example, that over a representative sample of ships' journeys, a proportion q of them result in a ship travelling to the stated destination D_s , then $P(H_0)$ could be set to q . In the absence of empirical evidence, this value will need to be set close to 1 (0.999 was chosen as a default value). This implies that $P(H_1)$ should be set at $1-q$. Since H_1 includes every destination that is not the stated one this covers $n-1$ destinations, where n is the total number of possible destinations. Hence the probability that the ship is travelling to a specific non-stated destination is $(1-q)/(n-1)$.

When the value of A_t falls below a threshold value, which will have to be set at a level that produces an acceptable number of false alarms, the ship's route is flagged up as being anomalous conditional on its expected destination being D_s . For the case where there is no knowledge of the expected destination, the use of this anomaly statistic can be extended to a series of null hypotheses for each possible destination D_i to determine at each observation, which of the destinations appear feasible and which appear anomalous. At the start of a journey all destinations will appear feasible (i.e. the $A_{t,i}$ value for each hypothesised destination D_i will be close to 1) but they will be gradually whittled down as the journey progresses and the route becomes clearer. If at any point of the journey, every possible destination has appeared anomalous at some previous point of the journey (including the current point) then there remains no feasible place for the ship to travel to and it should be flagged up as anomalous. In other words, $A_{t,i}$ has fallen below a defined threshold for every destination D_i at some time t (t is not constrained to be the same for each destination).

Analysis of the $A_{t,i}$ can also be useful in detecting where a change in apparent destination occurs.

5. Simple Network Example

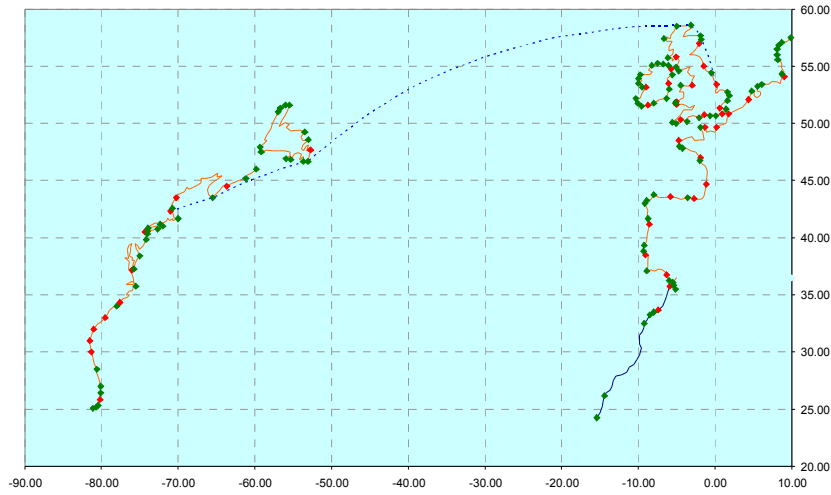


FIGURE 4. Nodes for a simple North Atlantic network with a journey from Boston to Grimsby shown

Figure 4 shows the nodes of a simple network based on the North Atlantic, which have been generated by the methods described in this paper. The lat/long co-ordinates have been represented in Cartesian form. The route shown (from Boston to Grimsby) takes the shortest distance between the two ports; in all this journey traverses 5 branches. Fig 5 shows a number of metrics associated with this journey.

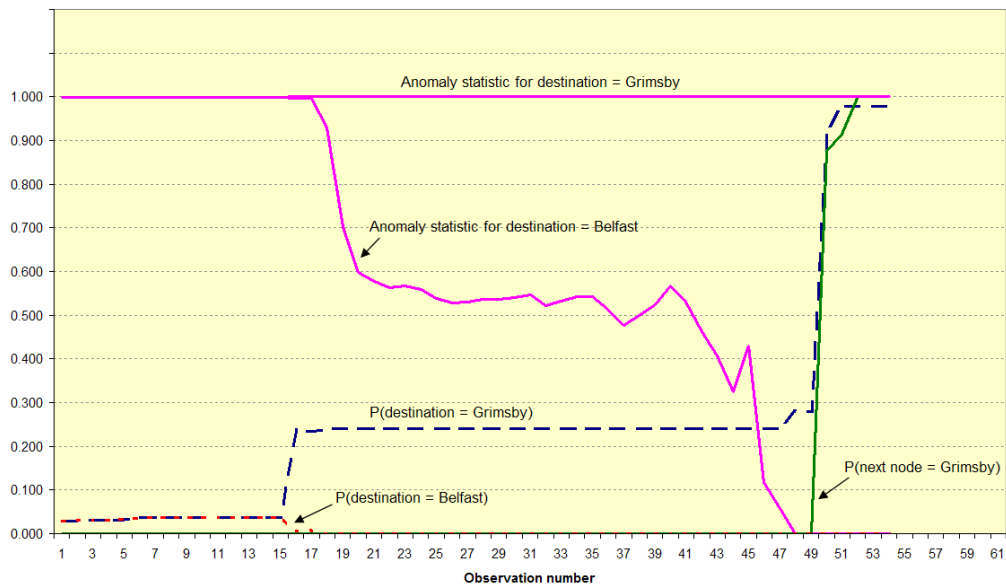


FIGURE 5. Metrics for the Boston to Grimsby route

The probability of the destination being Grimsby (dashed line) builds up incrementally at each node passed, finally leaping to almost 1 on the last branch. The anomaly statistic based on Grimsby being the destination remains at 1 throughout the journey, confirming that there has been no anomalous kinematic behaviour. However, the equivalent statistic based on the expected destination being Belfast falls sharply once the North American coast is passed and falls gradually to zero (the variation in the line reflects minor variations in the ship's headings). This would quickly be identified as being anomalous once it started to fall below 1. The dotted line (almost concurrent with the dashed line) is the probability of the ship's destination being Belfast which is also resolved and falls to zero once the ship has sailed past the American coast. The final solid line which jumps from 0 to 1 near the end of the journey is the probability of the next node being Grimsby.

5. Conclusions

Methodologies have been successfully developed for monitoring the global movement of ships and detecting whether or not they exhibit certain aspects of kinematic anomaly. The approach has been based around the development of a node-sparse network specifically suitable for ocean traffic. A Bayesian approach has been adopted for estimating the likelihoods for a range of journey destinations, upon which the anomaly statistics are based. The methodology as described can be substantiated by having access to individual ship historical data, but does not depend upon it.

REFERENCES

- (1) Lin Liao, Donald J. Patterson, Dieter Fox, and Henry Kautz. Learning and inferring transportation routines. *Artificial Intelligence*, 171(5-6):311-331, April 2007.(2)
- (2) National Geophysical Data Center (NGDC) coastline extractor: <http://rimmer.ngdc.noaa.gov>