UNIVERSIDAD DE LA LAGUNA

ESCUELA DE DOCTORADO Y ESTUDIOS DE POSGRADO

Detección de derivas en tráfico marítimo

Drifting Detection in Maritime Traffic

Autor:
Juan Manuel Falcón Ramírez

Tutores:
Dña. María Belén Melián Batista
D. José Marcos Moreno Vega

MÁSTER UNIVERSITARIO
EN CIBERSEGURIDAD E INTELIGENCIA DE DATOS

CURSO 2023 - 2024

Dña. **María Belén Melián Batista**, con N.I.F. 44311040E, Catedrática de Universidad adscrita al Departamento de Ingeniería Informática y de Sistemas de la Universidad de La Laguna, como tutora.

D. **José Marcos Moreno Vega**, con N.I.F. 42841047M, Catedrático de Universidad adscrito al Departamento de Ingeniería Informática y de Sistemas de la Universidad de La Laguna, como cotutor.

C E R T I F I C A N

Que la presente memoria titulada:

"*Detección de derivas en tráfico marítimo*"

ha sido realizada bajo su dirección por D. **Juan Manuel Falcón Ramírez**, con N.I.F. 45354218N.

Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en La Laguna a 26 de mayo de 2025.

Agradecimientos

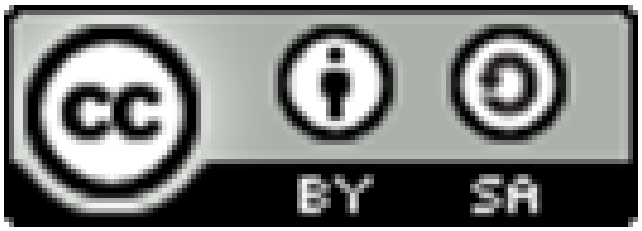
Me gustaría expresar mi más sincero agradecimiento a todas las personas que han contribuido a la realización de este trabajo.

Por un lado, agradezco a familia, amigos y compañeros. En especial, me gustaría agradecer a Cecilia por prestarme su apoyo incondicional, a Eloi por hacer el día a día más llevadero, y al grupito de amigos de la columna derecha con los que he sobrevivido al máster.

Por otro lado, quiero dar las gracias a aquellos profesores que han mostrado su apoyo al alumnado y han impartido clases de calidad. En concreto, me gustaría agradecer a mis tutores del proyecto, Marcos y Belén, por mostrarse dispuestos a ayudar en todo momento y llevar un seguimiento continuo del trabajo realizado.

Gracias a todos por estar ahí para hacer de este año una experiencia enriquecedora y amena.

Licencia



Resumen

El Sistema de Identificación Automática (AIS) es crucial para la navegación marítima moderna, proporcionando datos detallados sobre la posición, rumbo y velocidad de los barcos realmente útiles para la mejora constante de los sistemas de predicción de rutas y, en consecuencia, para mejorar la logística y la regulación del tráfico marítimo. El siguiente estudio se centrará en hacer uso de llamadas AIS realizadas por una serie de barcos en EE.UU., con el objetivo de detectar derivas intencionadas en las trayectorias de los barcos antes de su llegada a puerto, utilizando un algoritmo de K-vecinos más cercanos. La metodología consiste en la creación de matrices de probabilidad para distintos horizontes temporales (15, 30 y 60 minutos). Los resultados obtenidos alcanzan una precisión global de 0.85 en el mejor de los casos.

Abstract

The Automatic Identification System (AIS) is crucial for modern maritime navigation, providing detailed data on the position, heading, and speed of ships, which is really useful for the constant improvement of route prediction systems and, consequently, for improving the logistics and regulation of maritime traffic. The following study will focus on making use of AIS calls made by a number of ships in the USA, with the aim of detecting intentional drifts in ship trajectories prior to port arrival, using a K-nearest neighbours algorithm. The methodology is based on the creation of probability matrices for different time horizons (15, 30 and 60 minutes). The results obtained reach an overall accuracy of 0.85 in the best case.

Palabras clave: AIS, predicción, drifting, deriva, ruta, celda, algoritmo

Keywords: AIS, prediction, drifting, drift, route, cell, algorithm

Índice

Índice de figuras

Índice de cuadros

1 Introducción

Introduction

Navigation systems for transport have advanced significantly, especially in maritime traffic, with the use of the Automatic Identification System (AIS). This system, designed to avoid collisions, allows the exchange of navigation information between ships and shore stations, including data such as position, heading, speed and type of cargo.

Trajectory analysis and prediction are crucial to improving these navigation systems. An essential factor is drifting, which can be described as unexpected variations in a ship's trajectory due to safety reasons, weather conditions or other intentions. This study focuses on one type of intentional drifting, commonly used to manage port arrival time and avoid congestion: ships may choose to drift to adjust their arrival if the port is unavailable or undergoing maintenance.

Los sistemas de navegación para los medios de transporte han mejorado drásticamente con el paso de los años. En el caso del tráfico marítimo, la tecnología que se utiliza por excelencia es el sistema de identificación automática o AIS, por sus siglas en inglés (Automatic Identification System). Fue inicialmente concebido para evitar colisiones, permitiendo el intercambio de información de navegación entre barcos y estaciones en tierra. Los mensajes AIS contienen información como la posición (latitud y longitud), rumbo, velocidad y otros datos adicionales, como el tipo de barco o la carga que lleva (véase la tabla 1). Como cabría esperar, intercambios constantes de información dan lugar a una ingente cantidad de datos, de los cuales puede extraerse información de todo tipo.

Variables de las llamadas AIS	
Variable	**Descripción**
`BaseDateTime`	Fecha y hora en que se tomaron los datos (formato YYYY-MM-DD hh:mm:ss)
`Geolocalization`	Coordenadas de longitud y latitud del barco
`Heading`	Rumbo (orientación de la proa)
`COG`	Rumbo sobre el fondo (dirección real de movimiento del barco)
`SOG`	Velocidad sobre el fondo (en millas náuticas)
`Status`	Estado de movimiento del barco
`Vessel Type`	Tipo de embarcación
`Cargo`	Tipo de mercancía
`Ship dimensions`	Longitud, anchura y calado del barco

Cuadro 1: Información provista por las llamadas AIS.

Por la naturaleza de los datos AIS, el análisis y la predicción de trayectorias resultan de vital importancia para la continua mejora de los sistemas de navegación. Uno de los detalles a tener en cuenta durante el entrenamiento de los algoritmos de predicción es el concepto de deriva, o 'drifting' en inglés, que consiste en variaciones no esperadas en la trayectoria de un barco. Estas pueden deberse a motivos de seguridad, como evitar un obstáculo u otro barco, a situaciones climáticas adversas que obligan a cambiar el rumbo y la trayectoria de navegación, etc. Por otra parte, el drifting puede ser intencionado: existen razones científicas, como dejar el barco a la deriva para el estudio de las corrientes océanicas; o más mundanas, como empezar una deriva para abarcar más area durante la pesca.

De entre todas las posibilidades, el foco de estudio de este trabajo será un tipo de drifting más común e intencionado. Para una correcta organización del tráfico marítimo, a los barcos se les asigna unas ventanas de tiempo para poder atracar en puerto, con el fin de evitar mareas o congestión en la entrada de éste. Si un barco llega demasiado temprano, puede no haber un espacio disponible para él. Si el puerto está ocupado o en mantenimiento, es posible que no pueda recibir al barco a la hora prevista. En tales casos, los barcos pueden optar por derivar intencionalmente para ajustar su tiempo de llegada al puerto, evitando así pagar la estancia en un fondeadero.

Este tipo de drifting es problemático, dado que puede causar congestión en las áreas cercanas al puerto, dificultando la gestión del tráfico marítimo y aumentando el riesgo de accidentes. Dificultar la gestión del tráfico marítimo también conlleva repercusiones económicas significativas: los retrasos en la entrada y salida de los barcos pueden resultar en mayores costos operativos y pérdidas para las compañías navieras y los operadores portuarios. Además, la incertidumbre en los tiempos de llegada puede afectar a las cadenas de suministro, causando demoras en la entrega de mercancías y afectando a múltiples industrias que dependen del transporte marítimo. Por ello, resulta evidente la importancia de desarrollar algoritmos que puedan predecir y manejar estas derivas.

Este trabajo esta dividido en capítulos, con la finalidad de exponer de forma intuitiva las distintas fases de realización del proyecto. En primer lugar, en el capítulo 2 se explicarán los objetivos propuestos. En el capítulo 3, se expondrá la metodología aplicada y los motivos de ésta. El capítulo 4 contiene el análisis y procesado que se ha realizado previamente sobre los datos, con el fin de obtener un conjunto apropiado para el entrenamiento y validación del algoritmo a crear. En el capítulo 5, se estudia el criterio escogido para determinar qué considerar como drifting, además del diseño y la implementación del algoritmo sobre los datos. Los resultados obtenidos para cada caso de estudio y una breve discucón sobre ellos se muestran en la sección 6, seguidos por una conclusiones finales y posibles líneas futuras del proyecto comentados en el capítulo 7. Finalmente, se sintetizan todos los contenidos del proyecto en el capítulo 8.

2 Objetivos

Aims

The main objective of this project is to detect drifts in the trajectories of ships prior

to their arrival in port. As a secondary objective, the aim is to identify the starting point of the drift. For this purpose, a prediction algorithm based on the k-nearest neighbour classifier is proposed. The applications of this algorithm include improved predictions and the implementation of optimised trajectories to avoid intentional drifts, as well as improving logistics coordination and maritime traffic regulation.

El objetivo principal del proyecto consiste en detectar derivas en las trayectorias de los barcos en instantes previos de su llegada a puerto. Como objetivo secundario, se plantea detectar el punto de inicio de la deriva. Con esta finalidad, se propone crear un algoritmo de predicción basado en un clasificador de k-vecinos más cercanos, siguiendo la metodología especificada en Lo Duca et al. (2017). Las aplicaciones de un algoritmo con resultados aceptables son variadas, pero las más directas son las mejoras en las predicciones y la implementación de nuevas trayectorias optimizadas para evitar derivas intencionadas, mejorando así la coordinación de logística (reducción de esperas en los puertos) y la regulación del tráfico marítimo.

3 Metodología

Methodology

The methodology is based on Lo Duca et al. (2017), where the use of a K-NN classifier is suggested for drift detection in ship trajectories. Although subsequent research has explored other classifiers, such as Naive Bayes and support vector machines, k-NN and decision trees remain the best options (Lo Duca and Marchetti, 2020). More advanced methods, such as Osekowska et al. (2017) potential fields for more accurate drifting detection, will not be used in this work.

To train the model, a mesh will be superimposed on the ship's trajectory before reaching port, covering a two-hour stretch. A K-NN algorithm will be used to create a probability matrix indicating the likelihood of the ship being at a specific location. This matrix will be generated for three δ time increments: short (15 minutes), medium (30 minutes) and long term (60 minutes), thus allowing different types of drift to be identified.

La metodología seguida en el presente trabajo sigue las etapas fundamentales de la extracción de conocimiento desde bases de datos. En primer lugar, se recopilan y preparan los datos, destacando la necesidad de identificar rutas en ellos, con la intención de diseñar un algoritmo de aprendizaje capaz de detectar aquellas rutas que desarrollan drifting. Posteriormente, se ajustan experimentalmente los parámetros del algoritmo para optimizar su rendimiento. Luego, se valida el modelo en un entorno controlado y, finalmente, se extraen conclusiones significativas acerca del experimento y las bases del drifting.

Los conceptos teóricos se basan en la metodología llevada a cabo por Lo Duca et al. (2017). En el artículo de Lo Duca and Marchetti (2020), que continúa la investigación iniciada en el artículo anterior de los mismos autores, se proponen nuevos algoritmos de predicción basados en métodos como el Naive Bayes o máquinas de vectores soporte, pero el algoritmo k-NN sigue resultando ser la mejor opción (junto a los árboles de decisión). Existen métodologías más intensivas, como la propuesta por Osekowska et al. (2017), con la implementación de campos de potencial que afinan la detección del momento en el que empieza y termina el drifting, pero no serán utilizadas en este trabajo.

Imitando los pasos seguidos por Lo Duca et al. (2017), para el entrenamiento del modelo se creará una malla que se superpondrá sobre la trayectoria del barco. En el caso de este proyecto, se estudiarán tramos de dos horas justo antes de llegar a puerto en lugar de tramos en mar abierto. Con un algoritmo de vecinos más cercanos, se creará una matriz de probabilidades que responde a la pregunta "¿qué probabilidad hay de que el barco se encuentre en la posición x?", correspondiéndole una casilla de la malla a cada elemento de la matriz. Se creará una matriz para tres incrementos de tiempo: a corto, 15 minutos; a medio, 30 minutos; y a largo plazo, 60 minutos; pudiéndose así diferenciar distintos tipos de deriva.

4 Análisis y limpieza de los datos

Analysis and pre-processing of data

The project focuses on detecting drifting in ship trajectories prior to port arrival, using a K-nearest-neighbour classifier. Of all the features provided by the AIS calls, only longitude and latitude coordinates, the variable of the time at which the AIS call was made (in order to calculate the short, medium and long term position) and the 'Status' variable, which indicates whether the ship is grounded or moving, will be used. Other variables, such as speed or heading, will be discarded.

For the analysis, 170.000 AIS calls from 19 ships between various US ports were used. Incomplete routes were completed with synthetic data, resulting in a set of 17.425 AIS calls on 275 routes. To improve the training set, only the Miami arrival routes were selected, reducing the set to 9,751 calls on 130 routes.

During pre-processing, routes were identified using the ship's 'Status' and completed with synthetic data where necessary. From the complete dataset, routes consisting of several scattered points not showing a clear trajectory were eliminated. Then, routes were manually checked to confirm the presence of drifting, identifying 111 common trajectories and 19 with some drifting.

To increase the effectiveness of the model, synthetic trajectories with drifting were generated from original routes without drifting. Furthermore, synthetic trajectories without drifting were made, with the intention of improving the number of near neighbours and hence the accuracy of the model.

4.1. Variables relevantes

De las características proporcionadas por las llamadas AIS, solo se emplearán las coordenadas de longitud y latitud, la variable del momento en el que se hizo la llamada AIS, para poder calcular la posición a corto, medio y largo plazo, y la variable 'Status', que indica si el barco se encuentra varado o en movimiento. Esta última es relevante a la hora de diferenciar las distintas travesías cursadas por cada barco. A diferencia de lo que ocurre tanto en Lo Duca et al. (2017) como en Osekowska et al. (2017), que se emplean la velocidad y la dirección del barco para el entrenamiento del modelo, en este caso se tratará de detectar el *drifting* únicamente basándonos en la trayectoria predicha por el modelo k-NN.

4.2. Análisis exploratorio

Los datos utilizados para el estudio son de tráfico marítimo estadounidense. Se parte de base con 170000 llamadas AIS, procedentes de las trayectorias realizadas por 19 barcos diferentes entre los puertos de Miami, New Orleans, Jacksonville, Savannah, Mobile, North Charleston, Freeport y Tampa. La Figura 1 muestra, a modo de ejemplo, las distintas rutas obtenidas a partir de las llamadas AIS de uno de los barcos. Cada punto de las líneas trazadas representa la ubicación en la que se realizó la llamada AIS.

En un primer vistazo, cabe destacar que todos los barcos tienen en común que parten o llegan al puerto de Miami, dándose el caso para todas las rutas. Se pueden apreciar rutas que no acaban en ninguna parte debido a la falta de llamadas AIS para continuarlas. Para estos casos, se crearán datos sintéticos para unir la ruta a algún puerto cercano, simulando así la llegada del barco a puerto. Con las rutas completas, en esencia se tienen dos tramos de llegada a puerto: el puerto de llegada natural y el puerto de salida. Sin embargo, dependiendo del puerto de salida, la trayectoria al puerto destino puede llegar a variar significativamente, lo cual se tendrá en cuenta a la hora detectar si he producido o no deriva.

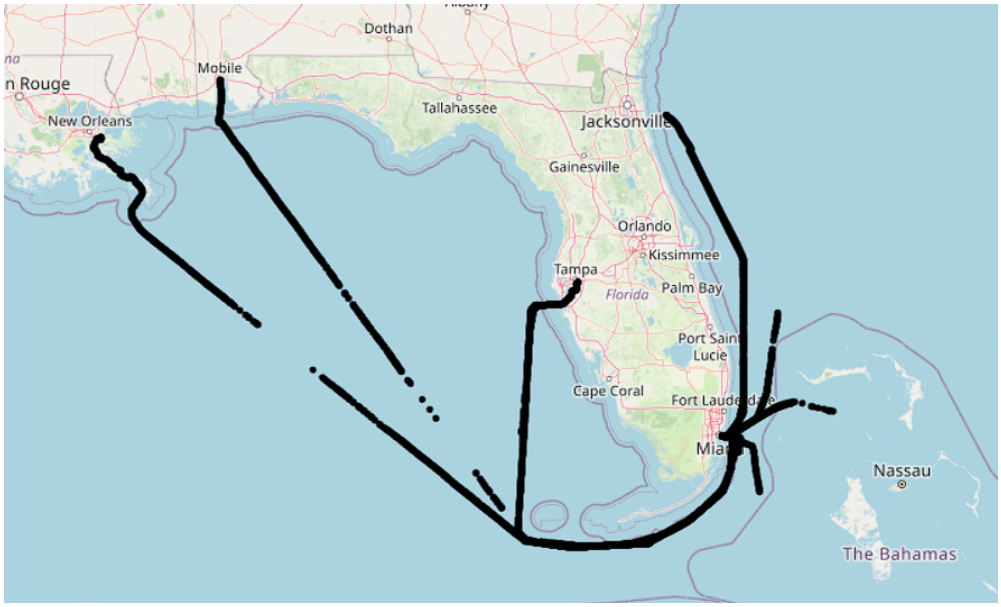


Figura 1: Ejemplo de rutas obtenidas a partir de las llamadas AIS.

4.3. Preprocesado

Para llevar a cabo la fase de preprocesado, en primer lugar, dentro del conjunto total de llamadas AIS se identifican las distintas rutas de cada barco, empleándose para ello la información que aporta la variable 'Status'. Un 0 en la variable indica que en el momento de la llamada el barco se encontraba en movimiento y con los motores en marcha, mientras que 1 indica que el barco estaba anclado, o un 5 que se encontraba varado en el puerto.

Del set completo de datos, se eliminan aquellas rutas que constan de varios puntos dispersos, es decir, aquellas que no muestran una trayectoria clara. Aquellas incompletas se rellenarán con datos sintéticos como se muestra en la Figura 2. Existen casos de trayectorias que no parecen llevar a ninguno de los puertos comunes y se acaban en alta mar. Estas rutas han sido completadas llevando el barco al puerto de Cancún o al de Cap-Haitien. A todas las rutas se les aplica una función que le asigna a cada punto una variable de longitud y latitud para un incremento de tiempo $\delta = 15$, 30 o 60 min. El resultado es un conjunto de 17.425 llamadas AIS repartidas entre 275 rutas diferentes.

Gran parte de las rutas estaban originalmente incompletas, por lo que una gran cantidad de llegadas a puerto se han rellanado sintéticamente, y cada una de estas llegadas parte de un punto inicial distinto (donde se cortaba la ruta original). Esto presenta un inconveniente a la hora de entrenar el modelo por vecinos más cercanos, dado que en múltiples casos los vecinos más cercanos se encuentran relativamente alejados entre sí y el programa los detecta como drifting, por el simple hecho de que ninguno de los puntos coincide con las casillas predichas por el modelo, a pesar de que el tramo en sí mismo es una línea recta que no presenta deriva ninguna. Para solucionar el problema, se ha propuesto utilizar únicamente las rutas de llegada a Miami, lo que vendría a ser aproximadamente la mitad del conjunto de datos después de eliminar las rutas inservibles. El problema sigue existiendo, pero en su mayoría se ve resuelto debido a la acumulación general de rutas en la zona cercana al puerto. Se cuenta entonces con 9751 llamadas AIS repartidas entre 130 rutas provenientes de los puertos anteriormente especificados.

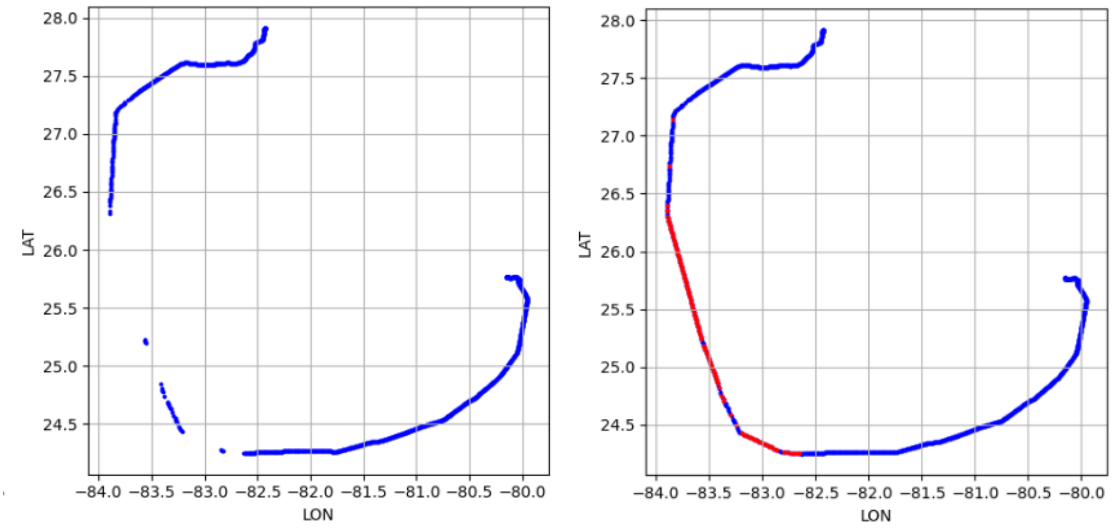


Figura 2: Ruta completada con datos sintéticos. En azul las llamadas AIS. A la derecha, en rojo, los datos sintéticos.

Con el objetivo de poder medir la efectividad del modelo, se han revisado a ojo cada una de las rutas anteriores para determinar si existe o no drifting en ellas. De las 130 totales, se ha considerado que 111 de ellas son trayectorias comunes y que las 19 restantes presentan drifting de algún tipo. En la Figura 3 puede apreciarse el conjunto al completo. Aunque no se muestre en el gráfico, a la hora de tratar con los datos se han eliminado las llamadas con coordenadas de longitud por debajo de -80.14, puesto que aproximadamente en ese punto comienza el puerto de Miami y las rutas siguen trayectorias fijas. El drifting solo debería aparecer previo a la llegada al puerto.

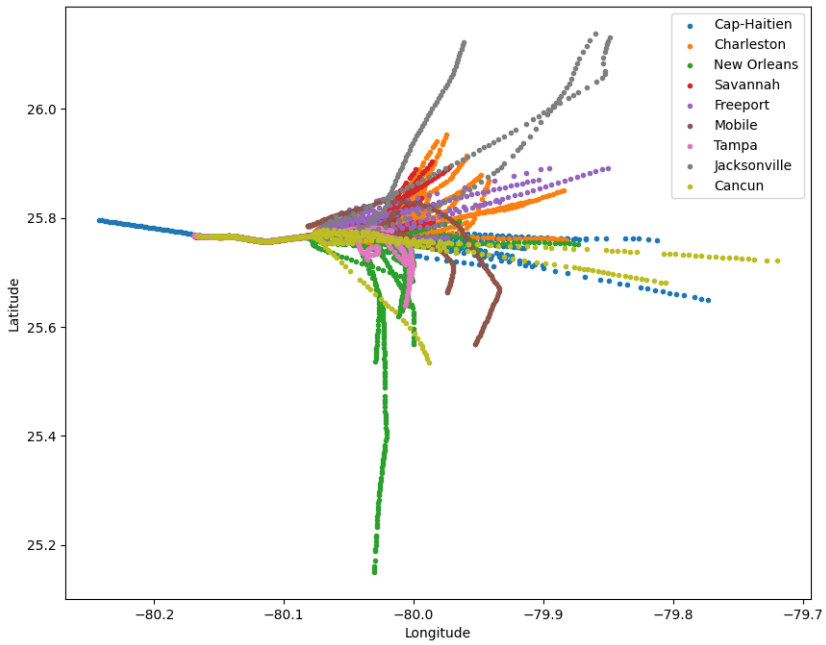


Figura 3: Tramo del puerto de Miami (2h). Se han pintado las trayectorias en función del puerto de destino/salida.

4.4. Creación de rutas sintéticas

Debido al limitado grupo de rutas originales que presentan drifting, de las cuales no todas presentan un drifting pronunciado, se ha planteado la posibilidad de generar rutas con deriva a partir de las rutas sin deriva ya existentes. Con un algoritmo aleatorio, los puntos de una de las rutas originales se verían modificados para crear una trayectoria con curvas o 'loopings', claros ejemplo de drifting intencionado. La Figura 4 muestra un ejemplo de ruta sintética generada aleatoriamente a partir de una de las rutas que llegaban de Savannah. En algunos casos, el resultado de estas deformaciones puede ser exagerado, creándose una trayectoria imposible para un barco real, por lo que la ruta se descartaría.

Utilizando el mismo algoritmo, modificando previamente los parámetros, se han generado también rutas completamente sintéticas sin deriva que siguen los patrones de trayectoria de las rutas originales. La idea detrás de esta tarea es la de contar con suficientes vecinos cercanos para cada una de las rutas con drifting, evitando así el problema de que se detecten derivas por el mero hecho de que las rutas más cercanas a la ruta estudiada se encuentren alejadas de ésta. En teoría, esto permitiría aumentar la efectividad

del modelo, puesto que se cuenta con una mayor cantidad de rutas para el entrenamiento.

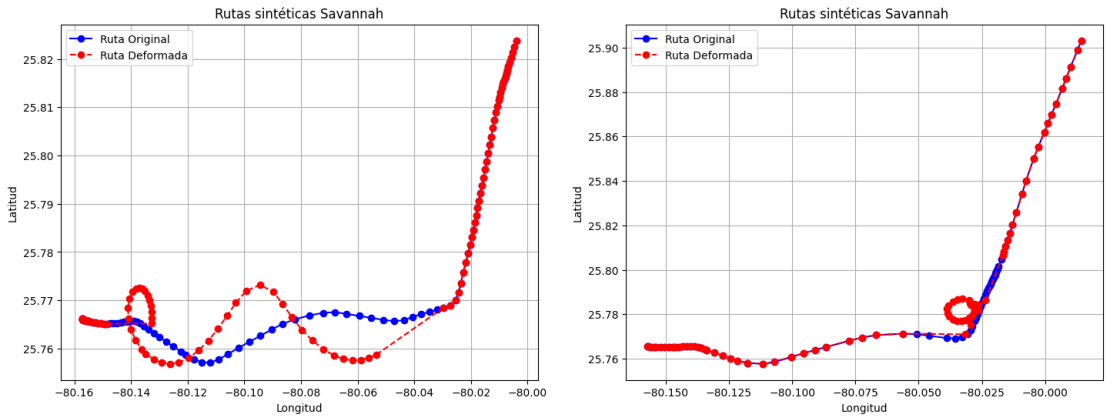


Figura 4: Ejemplos de rutas sintéticas. En azul, la ruta original. En rojo, la ruta creada a partir de la original.

5 Algoritmo de detección de drifting

Drifting detection algorithm

Following the methodology of Lo Duca et al. (2017), the position of the ships at δ 15, 30 and 60 minutes were calculated. The drift will be detected by creating a grid overlaid over the routes. The number of cells per dimension n varies according to the length of the routes: long routes require fewer cells, and short routes require more. k-NN models are trained with probability matrices for each coordinate and time δ, and each cell of the grid corresponds to an element of the matrix. The cell with the highest probability indicates the predicted position of the ship at time δ. Cells with higher probability are shown in vivid colours and the original trajectories in attenuated pink to identify errors in drifting detection.

The algorithm's criterion for detecting drift is based on the K-NN model classification. For each time interval δ, the predicted cells are calculated. If a point of the route is not in a cell with more than 80 % probability, it is marked. If five or more consecutive points are outside the predicted cells, drift is considered, and it starts at the first point marked. If, before reaching five points out of place, one falls into a predicted cell, the list is cleared and no drift is detected. Drift is considered to be present if it is detected in any of the three time intervals δ.

5.1. Creación de las celdas de predicción

La estructura del algoritmo utilizado para la detección del drifting se observa en el diagrama de la Figura 5. Siguiendo los principios de la metodología de Lo Duca et al. (2017) y habiendo calculado la posición dentro de $\delta = 15$, 30 y 60 minutos para cada una de las llamadas AIS, se comienza el algoritmo de predicción con la creación de una malla superpuesta sobre las rutas. Dicha malla contará con un parámetro variable: el número de celdas por dimensión n, el cual tendrá especial relevancia a la hora de calcular las predicciones. Dado que los puertos de procedencia son diferentes, en el cálculo de las dos horas antes de llegar a puerto algunas rutas son más largas que otras, como se pudo ver en la Figura 3. En las rutas más largas se obtienen buenas predicciones con números de celda bajos, mientras que para las rutas más cortas son necesarias muchas más celdas. Esto se debe principalmente al hecho de que la malla tiene el mismo tamaño para cada caso, basado en la latitud y longitud máxima encontrada en las rutas del conjunto de que se utilice para el entrenamiento. Un ejemplo puede apreciarse en la Figura 6, que muestra una ruta de corta longitud con deriva. En este caso, se necesitan celdas más pequeñas para poder detectar el drifting, lo que implica aumentar el número n.

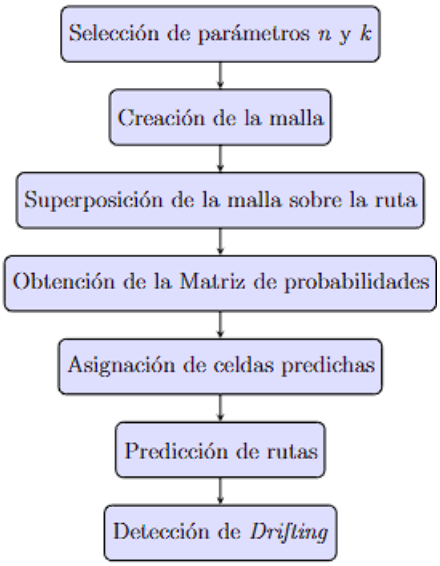


Figura 5: Diagrama de pseudocódigo.

A continuación, se entrenan los modelos k-NN de clasificación, calculándole a cada coordenada de los distintos δ de tiempo una matriz de probabilidad. Cada elemento de esta matriz hace referencia a una de las celdas de la malla. Las probabilidades se obtienen en base a las k rutas más cercanas: para cada llamada AIS de las rutas que comprenden los k vecinos, se calcula la posición del barco a un tiempo δ y se anota en qué celda de la malla se ubicaría. Suponiendo $k = 5$ vecinos, cada celda anotada sumará un 20 % de probabilidad, llegando al 100 % si las 5 rutas vecinas coinciden en que el barco acabará en la misma celda. Una vez aplicado el algoritmo, la celda que tenga mayor probabilidad se le asignará a una nueva variable *grid-*δ agregada a la llamada AIS correspondiente, y el trazado de esta variable para todas las llamadas AIS que componen la ruta indica la trayectoria predicha para el barco. Al representar la ruta junto a la malla y las celdas predichas, se verán con un color más intenso aquellas con mayor probabilidad y con un

color más tenue las que tienen menor probabilidad (véase la Figura 6). En un color rosado atenuado se muestran las trayectorias seguidas por las rutas originales, principalmente para poder comprobar en qué casos se detecta erróneamente el drifting debido a falta de rutas cercanas.

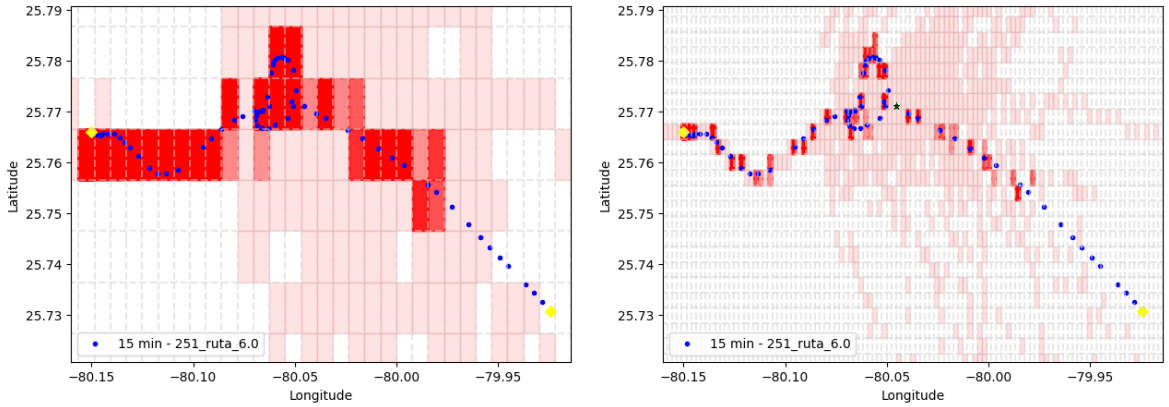


Figura 6: Malla de probabilidades. A la izquierda, 50 celdas por dimensión, no se detecta drifting. A la derecha, 200 celdas por dimensión, sí se detecta drifting (la estrella negra marca el inicio de este).

5.2. Criterio para la detección

El criterio empleado por el algoritmo para detectar drifting se basa en la clasificación realizada por el modelo k-NN. Para cada δ, se calculan las celdas predichas. Se especifican dos nuevos parámetros: un umbral de probabilidad z_{prob}, que determina el mínimo de probabilidad necesario en una celda determinada para considerar que el barco se encuentra en la ruta predicha; y un mínimo de puntos consecutivos pc_{min}, que establece la cantidad necesaria de llamadas AIS seguidas que se desvían de la ruta predicha para considerar que se ha producido drifting. Se ha encontrado, tras reiteradas aplicaciones del algoritmo, que los mejores valores de estos parámetros —para el conjunto de datos estudiado— son $z_{prob} = 0.8$ y $pc_{min} = 5$. Con un umbral de probabilidad del 80 % se garantiza que la celda que se predice para la llamada AIS forma parte de la ruta predicha por el modelo, además de que la celda quedaría respaldada por la mayoría de rutas que componen los k vecinos, evitándose casos donde dos o más celdas compartan la probabilidad más alta y sólo una de ellas se ha escogido para representar la ruta predicha. Por otra parte, $pc_{min} = 5$ es un buen punto intermedio, dado que no se necesitan demasiados puntos para considerar drifting, por lo que las derivas cortas serán detectadas, ni se necesitan muy pocos, lo que haría que se detectasen derivas donde no las hay.

A modo de ejemplo de cómo el modelo aplica el criterio de detección: si un punto de la ruta estudiada a tiempo δ no está en contacto con una de las celdas predichas de más de 80 % de probabilidad, se anota la posición del punto y se le añade a una lista vacía. Si el siguiente punto tampoco se ajusta a las predicciones, se añade a la lista. Si en algún momento la lista llega a tener 5 o más puntos consecutivos fuera de lugar, se considera que existe deriva y que comienza en el primer punto anotado. En el caso de que, antes de que la lista llegue a 5 puntos, se encuentre un punto en una casilla predicha, se borra la lista y se entiende que no hubo drifting. De este modo, se tienen en cuenta pequeñas

desviaciones. Por último, aunque no se trate de un suceso habitual, destacar que existe la posibilidad de que el drifting únicamente aparezca para uno de los tres periodos de tiempo considerados (15, 30 o 60 minutos). En estas situaciones, se determinará que la ruta tiene drifting desde que se detecte para cualquiera de los tres casos.

6 Resultados y discusión

Results and discussion

In the first scenario, the set of 130 original routes was used, of which 111 were drift-free. This set is divided into 80 % for training and 20 % for testing, mixing the routes with and without drift in the test group. A grid approach is used to determine the best cell parameters by dimension and nearest neighbours, finding that the number of cells per dimension $n = 100$ gives the best, but poor results due to the lack of routes in the training set. In the second scenario, all original non-drift routes were used for training, and the test set includes both drift and non-drift synthetic routes. This results in a substantial improvement in the accuracy of the model, notably better performance in drift-free route detection, which makes increasing n a good option. The best fit is found at $n = 200$, suggesting that a larger training set similar to the test set leads to a better predictive capability of the algorithm.

The results show that cell size is critical to detecting drifts in maritime routes and that the number of nearest neighbours, k, has no influence. Long routes tend to be detected as drifts due to the lack of close neighbours in the grid, while short routes may go undetected. A new function was proposed in order to improve drift detection: a distance to the port of Miami of 10 km instead of the time of 2h is used to obtain the routes, resulting in a more compact dataset of 126 routes without drift and five routes with it.

Again, two data sets were evaluated: one with the original routes and one combining original data with synthetic data. Using a grid to determine the best parameters, it was found for the first case that $n = 30$ gave the best results, with a higher recall compared to routes taken two hours from the port. However, specificity indicated that drifts were still detected on routes with no drift, resulting in low positive accuracy. On the other hand, when combining original and synthetic data, $n = 100$ achieves the best statistics. Relative to the time-based dataset, a slight decrease in recall and higher specificity was observed, with an overall improved accuracy. The various case studies show that a training set covering all common port arrival routes, including small variations to take into account ships deviating slightly above or below the planned routes, is necessary to obtain good results.

El planteamiento seguido para la obtención de unos resultados justificables y de calidad se basa en la evaluación previa de los escenarios de estudio contemplados. Se parte de la base de que las rutas identificadas cuentan con tamaños muy diferentes, por lo que el parámetro ajustable n cobra mayor peso. Se han realizado experimentos ajustando

el modelo para distintos valores de n, en múltiplos de 10, hasta encontrar un balance equilibrado donde el algoritmo sea capaz de detectar drifting tanto en rutas largas como en cortas. Lo mismo se aplica a la variable k. Se propone entonces un segundo escenario con rutas más similares en tamaño, con la finalidad de comprobar si es posible reducir la influencia de n y que los datos no se vean sesgados por la determinación de este parámetro.

Por otro lado, existe el problema de que el conjunto de datos de entrenamiento no sea lo suficientemente grande como para abarcar rutas prácticamente iguales en trayectoria, pero ligeramene desplazadas en posición. Se proponen dos casos para cada escenario: uno que utiliza los datos originales para entrenamiento y validación; y otro que reserva los datos originales sin deriva únicamente para el entrenamiento, añadiendo datos con y sin deriva al conjunto de testeo. En ambos casos y ambos escenarios, se analizará el promedio de las distintas medidas estadísticas evaluadas por el modelo tras realizarse 15 iteraciones de éste sobre 15 conjuntos de testeo diferentes seleccionados aleatoriamente. En los casos donde sólo se utilizan los datos originales, el conjunto de entrenamiento también variará de forma aleatoria, estableciendo su tamaño en el 80 % del total de los datos sin deriva.

6.1. Tramos de dos horas antes de llegar al puerto

6.1.1. Conjunto de rutas originales

Como primer escenario, se considerará el conjunto original que cuenta con 130 rutas. Dentro de las 111 que no presentan deriva, se ha escogido un grupo de entrenamiento con el 80 % de las rutas, y el 20 % restante, 33 rutas en total, para el testeo. El 100 % del conjunto de entrenamiento no presenta drifting, puesto que es la única forma de entrenar al algoritmo correctamente. Para el grupo de testeo, se mezclarán las 19 trayectorias con deriva con las 33 que no tienen.

Para determinar el mejor número de celdas por dimensión n y el mejor número de vecinos más cercanos k, se ha creado un grid con puntuaciones ponderadas, que prioriza la precisión global. El primer detalle que se observa es que la cantidad de vecinos más cercanos no influye en el modelo. De 2 vecinos en adelante, ninguno de los parámetros estadísticos evaluados varía. Por otro lado, el número n está fuertemente correlacionado con la cantidad de derivas detectadas. Un número muy bajo apenas las detecta, mientras que un número muy alto detecta derivas donde no las hay. Atendiendo al grid, el mejor número de celdas por dimensión para un resultado equilibrado es de $n = 100$. No obstante, los resultados son bastante pobres, principalmente debido a que faltan rutas para un conjunto de entrenamiento más completo, donde cada ruta tenga unos vecinos más cercanos claros. Números más altos de n detectan mejor el drifting pero empiezan a perder precisión para la clase negativa rápidamente. La tabla 2 muestra los resultados promedios de haber ejecutado el código con 15 subconjuntos de testeo aleatorios distintos. $\hat{\text{Y}}$ y $\hat{\text{N}}$ hacen referencia a las categorías reales, mientras que Y y N muestran la clasificación por parte del algoritmo.

	N	Y	TOTAL
$\hat{\text{N}}$	26.3	6.7	33
$\hat{\text{Y}}$	10.8	8.3	19
TOTAL	∼37	15	52

Medidas promedio	
Accuracy	0.67
Positive precision	0.55
Negative precision	0.71
Recall	0.42
Specificity	0.79

Cuadro 2: Tramo de 2h. Resultados del clasificador con las rutas originales ($n = 100$).

Podemos destacar de los resultados un recall bajo, lo que implica que el modelo no está identificando correctamente los casos de drifting reales, un resultado esperable teniendo en cuenta el problema que supone regular la variable n. Por otro lado, la especifidad es relativamente alta, lo que implica que, en general, no se están detectando casos de drifting de más. Observando la precisión, se puede concluir que el algoritmo en estas condiciones resulta de mucha utilidad para detectar drifting.

6.1.2. Conjunto de rutas originales y sintéticas

Para el segundo escenario, el 100 % de las rutas originales sin deriva serán utilizadas para el entrenamiento. El conjunto de testeo estará formado por dos subgrupos: el grupo con drifting, que mezcla aleatoriamente las 19 rutas con deriva originales con las creadas sintéticamente; y el grupo sin drifting, formado por rutas sintéticas creadas a partir de las originales que presentan pequeñas variaciones que no llegan a considerarse derivas. Como cabría esperar, un conjunto de entrenamiento más extenso que se asemeja al conjunto de testeo (rutas cercanas entre sí, lo esperable en un caso real) obtiene resultados considerablemente mejores, tal y como se muestra en la tabla 3. Las rutas que no tienen deriva se detectan con mayor facilidad, lo que permite aumentar n para encontrar las derivas sin necesidad de perder precisión en la otra clase. En este caso, el mejor número de celdas por dimensión sin perder en especificidad es $n = 200$.

	N	Y	TOTAL
$\hat{\text{N}}$	30.9	6.1	37
$\hat{\text{Y}}$	3.4	14.8	18
TOTAL	∼34	∼21	55

Medidas promedio	
Accuracy	0.83
Positive precision	0.71
Negative precision	0.90
Recall	0.83
Specificity	0.84

Cuadro 3: Tramo de 2h. Resultados del clasificador con el conjunto de rutas originales y sintéticas ($n = 200$).

6.2. Tramos de 10 km antes de llegar a puerto

Los resultados anteriores muestran la importancia del tamaño de las celdas a la hora de detectar derivas, pero esto ocurre en gran medida por la gran variedad de rutas presentes en el conjunto de entrenamiento. La realidad es que, al escoger las rutas basándonos en dos horas antes de la llegada como criterio, hay rutas mucho más largas que otras donde el algoritmo tiende a detectar que se ha producido drifting porque la mitad de la ruta no

tiene un vecino más cercano realmente cercano. El caso contrario también ocurre: rutas pequeñas con deriva pasan indetectadas, debido a que el grid está superpuesto sobre todas las rutas y, por ende, el tamaño de las casillas para esta clase de rutas es mayor, pudiendo caer varios puntos de la ruta en una misma celda predicha.

Por esta razón, con el objetivo de reducir la relevancia del parámetro n y obtener un mejor conjunto de datos a priori para el entrenamiento del modelo, con rutas más similares en tamaño y trayectoria, se les ha aplicado a las rutas originales una nueva función para obtener las celdas en tiempos δ, esta vez basada en la distancia con respecto al puerto de Miami. Se han recopilado las llamadas AIS de cada ruta en un radio de 10 km desde el puerto de Miami. El resultado de esta criba se aprecia en la Figura 7, con un conjunto de datos mucho más compacto independientemente del otro extremo de la ruta, salvo en casos excepcionales. El único inconveniente que conlleva este método es la pérdida de datos, relevante sobre todo a la hora de calcular la posición en $\delta = 60min$. No obstante, teniendo en cuenta que el foco de estudio es la detección de derivas en las cercanías al puerto, no debería suponer un problema centrar la atención en las derivas a $\delta = 15, 30$ min.

El conjunto completo tan sólo cuenta con cinco rutas con drifting, por lo que se completará con rutas sintéticas. Siguiendo el mismo principio que en el caso anterior, también se han creado rutas sintéticas sin drifting para poder realizar las mismas comparaciones. Otro detalle a tener en cuenta es que el tamaño de la malla de celdas ha disminuido como consecuencia de tener rutas más agrupadas, por lo que el número n puede verse reducido.

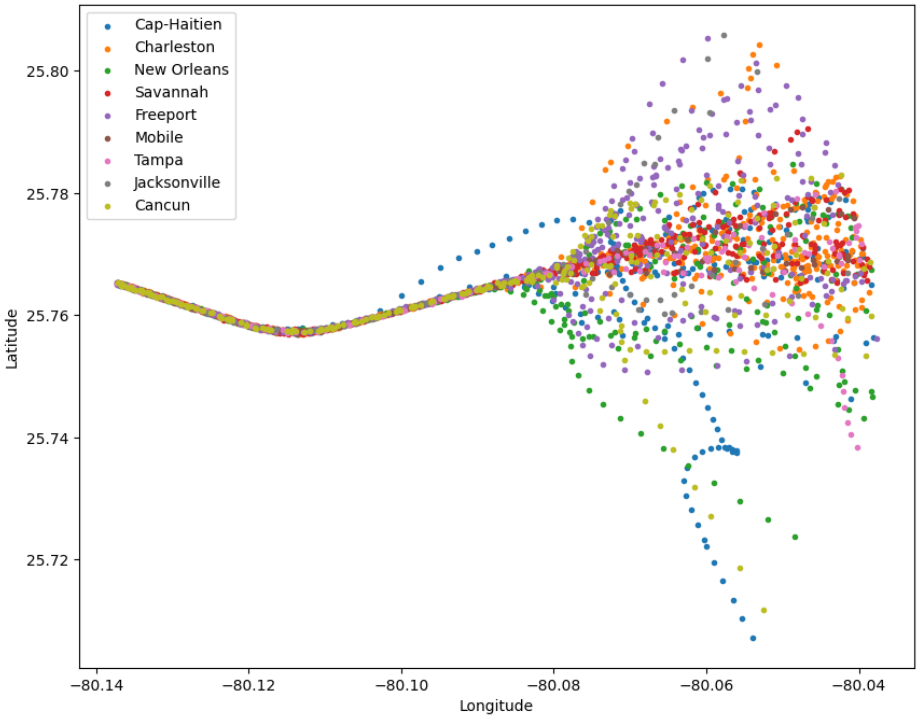


Figura 7: Tramo puerto de Miami (10 km de radio).

6.2.1. Conjunto de rutas originales

Con el grid para detectar los mejores parámetros, se llega a la conclusión de que utilizar $n = 30$ celdas por dimensión garantiza los mejores resultados. La tabla 4 muestra un recall un poco más alto en comparación con las rutas tomadas a dos horas del puerto, lo que indica que los casos reales se detectan mejor. La especificidad muestra que aún se detectan derivas en rutas sin drifting, lo que da lugar a un precisión positiva muy baja. En términos generales, los resultados son muy parecidos.

	N	Y	TOTAL
$\hat{\text{N}}$	28.7	8.3	37
$\hat{\text{Y}}$	8.9	9.1	18
TOTAL	∼38	∼17	55

Medidas promedio	
Accuracy	0.69
Positive precision	0.52
Negative precision	0.76
Recall	0.51
Specificity	0.78

Cuadro 4: Tramo de 10 km. Resultados del clasificador con el conjunto de rutas originales ($n = 30$).

6.2.2. Conjunto de rutas originales y sintéticas

Este último caso presenta unos resultados similares a su equivalente tomando el tramo de 2h. La principal diferencia se da en el recall y la especificidad. Los resultados de la tabla 5 muestran un recall ligeramente más bajo, lo que implica que el modelo detecta menos casos reales de drifting, pero una especificidad más alta, lo que indica que no se están detectando como drifting las trayectorias sin deriva.

	N	Y	TOTAL
$\hat{\text{N}}$	33.8	3.2	37
$\hat{\text{Y}}$	4	9.6	∼14
TOTAL	∼38	∼13	51

Medidas promedio	
Accuracy	0.85
Positive precision	0.75
Negative precision	0.89
Recall	0.73
Specificity	0.91

Cuadro 5: Tramo de 10 km. Resultados del clasificador con el conjunto de rutas originales y sintéticas ($n = 100$).

7 Conclusiones y posibles líneas futuras

Conclusions and future work

The different cases studied show that a training set covering all common port arrival routes is necessary, including small variations that take into account ships deviating slightly above or below the prescribed routes. Thus, increasing the n parameter considerably improves drift detection and the overall accuracy of the model. On the other hand,

it has been determined that the k number of nearest neighbours does not influence the results, mainly because most routes are plotted in the same way: ships navigate along similar longitude and latitude coordinates for each route, which means that the cells they pass through will turn out to be the same.

The algorithm proposed seems to give satisfactory results as long as the training set is complete, regardless of the size of the routes analysed. Being more rigorous, the case for the segment 10 km from the port is the closest to reality, given that if the routes are extended further, it will reach a point where it is very difficult to replicate all the small variations that arise in the routes throughout the ship's voyage, which will cause drifting to be detected due to a lack of nearby neighbours. In addition, the most worrying cases of drifting are the intentional drifts that occur in the vicinity of the port, which is best represented by the set of routes at 10 km.

As possible future lines of work, the first option is to extend and diversify the training set, to include routes with significant variations and corresponding to more than one port. Another way would be to study the optimisation of the parameters of the model, particularly n, in different scenarios and contexts. Similarly, it would also be useful to develop algorithms that detect intentional drifting near ports by analysing patterns in the final segments of the routes. Finally, validation of the algorithm in real case studies and comparison with other prediction models may be helpful to improve its effectiveness.

Los distintos casos estudiados muestran que es necesario un conjunto de entrenamiento que abarque todas las rutas comunes de llegada a puerto, incluyendo pequeñas variaciones que tengan en cuenta aquellos barcos desviados ligeramente más arriba o más abajo de las rutas pautadas. De esta manera, aumentar el parámetro n mejora considerablemente la detección de las derivas y la precisión general del modelo. Por otro lado, se ha determinado que el número k de vecinos más cercanos no influye en los resultados, debido principalmente a que la mayoría de rutas son trazadas de la misma manera: los barcos navegan por coordenadas de longitud y latitud similares para cada ruta, lo que significa que las celdas por las que pasan resultarán ser las mismas.

El algoritmo planteado parece dar resultados satisfactorios siempre y cuando el conjunto de entrenamiento esté completo, independientemente del tamaño de las rutas analizadas. Siendo más rigurosos, el caso para el tramo a 10 km del puerto es el más cercano a la realidad, dado que si se siguen extendiendo las rutas, llegará un momento donde sea muy complicado replicar todas las pequeñas variaciones que van surgiendo en estas a lo largo de la travesía del barco, lo que causará que se detecte drifting por falta de vecinos cercanos. Además, los casos de drifting más preocupantes son los intencionados que suceden en las cercanías a puerto, lo cual queda mejor representado con el conjunto compuesto por las rutas a 10 km.

Como posibles líneas futuras de trabajo, destacar como primera opción la ampliación y diversificación del conjunto de entrenamiento, incluyendo rutas con variaciones significativas y tramos correspondientes a más de un puerto. Otro camino a seguir sería estudiar la optimización de los parámetros del modelo, particularmente n, para diferentes escenarios y contextos. Del mismo modo, también resultaría provechoso desarrollar algoritmos que detecten drifting intencionado cerca de los puertos mediante análisis de patrones en los

tramos finales de las rutas. Por último, la validación del algoritmo en estudios de caso reales y la comparación con otros modelos de predicción pueden resultar de gran ayuda para mejorar su eficacia.

8 Síntesis

Summary

A study was carried out to detect near-port drifts using AIS data from ships in the USA. The different routes traced by each of the ships were extracted from the AIS data, grouped by port and completed with synthetic data if necessary. The routes were then selected for two case studies: segments of 2 hours before reaching the port of Miami, and segments of a radius of 10 km, taking the port as the central point. It has been determined on which routes drifting is present and the k-NN algorithm to be trained has been designed, which consists in placing a grid over the routes and trying to predict in which cell of the grid the ship will most likely be found in a period of k-delta = 15, 30 or 60 min. The model considers drifting to be present if a series of five or more consecutive AIS calls are outside their respective predicted cells, with the first point in the series considered to be the start of drift.

For the training of the model, displaced routes have been simulated with synthetic routes, obtained by slightly deforming the original routes. In the same way, synthetic routes with drifting have been created. The results obtained show that a training set covering all the most common routes and some slightly shifted routes is necessary, so that for each case there are truly close neighbours. It was found that the number of nearest neighbours k does not influence the results, while the number of cells per grid dimension n is particularly relevant when detecting drifting: a higher number n detects more cases of drifting, which can be a problem if there is an appreciable difference between the length of the different routes used for training. Overall, the model is able to detect drifts correctly, with an overall accuracy of 0.85.

Con el propósito de detectar derivas en las cercanías a puerto, se han estudiado las llamadas AIS realizadas por una serie de barcos que navegan en EEUU. De ellas se han extraído las distintas rutas trazadas por cada uno de los barcos, se han agrupado por puertos y se han completado con datos sintéticos los tramos incompletos. A continuación, se han seleccionado las rutas para dos casos de estudio: tramos de 2h antes de llegar al puerto de Miami y tramos de un radio de 10 km tomando el puerto como punto central. Se ha determinado en qué rutas hay drifting y se ha diseñado el algoritmo k-NN a entrenar, que consiste en superponer una malla sobre las rutas y tratar de predecir en qué celda de la malla se encontrará con mayor probabilidad el barco en un periodo $\delta = 15, 30$ o 60 min. El modelo considera la presencia de drifting si una serie de cinco o más llamadas AIS consecutivas se encuentran fuera de sus respectivas celdas predichas, considerándose el primer punto de la serie como el inicio de la deriva.

Para el entrenamiento del modelo, se han simulado rutas desplazadas con rutas sintéticas, obtenidas deformando ligeramente las rutas originales. Del mismo modo, se han creado rutas sintéticas con drifting. Los resultados obtenidos muestran que es necesario un conjunto de entrenamiento que abarque todas las rutas más comunes y algunas rutas desplazadas ligeramente, de modo que para cada caso existan vecinos verdaderamente cercanos. Se ha determinado que el número de vecinos más cercanos k no influye en los resultados, mientras que el número de celdas por dimensión de la malla n es especialmente relevante a la hora de detectar el drifting: un número más alto detecta más casos de deriva, pudiendo ser un problema si existe una diferencia apreciable entre la longitud de las distintas trayectorias empleadas para el entrenamiento. En términos generales, el modelo es capaz de detectar derivas correctamente, habiéndose alcanzado una precisión total de 0.85 en el mejor de los casos.

Referencias

Angelica Lo Duca and Andrea Marchetti. Exploiting multiclass classification algorithms for the prediction of ship routes: a study in the area of malta. *Journal of Systems and Information Technology*, ahead-of-print, 07 2020. doi: 10.1108/JSIT-10-2019-0212.

Angelica Lo Duca, Clara Bacciu, and Andrea Marchetti. A k-nearest neighbor classifier for ship route prediction. In *OCEANS 2017 - Aberdeen*, pages 1–6, 2017. doi: 10.1109/OCEANSE.2017.8084635.

Ewa Osekowska, Henric Johnson, and Bengt Carlsson. Maritime vessel traffic modeling in the context of concept drift. *Transportation Research Procedia*, 25:1457–1476, 2017. ISSN 2352-1465. doi: https://doi.org/10.1016/j.trpro.2017.05.173. URL https://www.sciencedirect.com/science/article/pii/S2352146517304660. World Conference on Transport Research - WCTR 2016 Shanghai. 10-15 July 2016.