



Centro Universitario de la Defensa en la Escuela Naval Militar

TRABAJO FIN DE GRADO

Generación y evaluación de datos AIS dinámicos sintéticos

Grado en Ingeniería Mecánica

ALUMNO: Mercedes Sánchez de Toca Rodríguez

DIRECTORES: Belén Barragáns Martínez

Pablo Sendín Raña

CURSO ACADÉMICO: 2025-2026

Universida_{de}Vigo



Centro Universitario de la Defensa en la Escuela Naval Militar

TRABAJO FIN DE GRADO

Generación y evaluación de datos AIS dinámicos sintéticos

Grado en Ingeniería Mecánica
Intensificación en Tecnología Naval
Cuerpo General

Universida_deVigo

RESUMEN

Este Trabajo de Fin de Grado se centra en la generación sintética de datos AIS (*Automatic Identification System*) dinámicos mediante el uso de técnicas de aprendizaje automático, con el objetivo de modelar y reproducir trayectorias marítimas realistas. Este problema resulta especialmente relevante en contextos donde los datos disponibles son limitados, presentan desbalanceo o están sujetos a restricciones de acceso.

Para ello, se emplean modelos generativos basados en redes neuronales, concretamente arquitecturas LSTM y modelos variacionales como SeqVAE, diseñados para capturar la naturaleza secuencial y la dependencia temporal de las trayectorias. La generación de datos se realiza con modelos entrenados con trayectorias reales, buscando preservar tanto la estructura temporal como las características espaciales del movimiento de los buques.

Los datos sintéticos generados se evalúan mediante un enfoque combinado que incluye un análisis estadístico de las características de las trayectorias generadas y, especialmente, inspección visual de estas sobre mapas. Este enfoque permite analizar su coherencia espacial, escala, distribución geográfica y variabilidad. Asimismo, se estudia el comportamiento de los modelos en función del tipo de buque, identificando diferencias en la capacidad de reproducción de los patrones de navegación.

Los resultados muestran que los modelos son capaces de capturar el comportamiento global de las trayectorias, aunque presentan limitaciones en la representación de la variabilidad y en la fidelidad de la distribución espacial. Esto evidencia tanto el potencial de estos enfoques como la necesidad de continuar mejorando su capacidad generativa.

PALABRAS CLAVE

Datos sintéticos, AIS, trayectorias, LSTM, SeqVAE

AGRADECIMIENTOS

Quiero expresar mi más sincero agradecimiento a todas las personas que han formado parte de este camino y que, de una u otra manera, han contribuido a la realización de este Trabajo de Fin de Grado. Contribuyendo no solamente a este proyecto, sino también al cierre de una de las etapas más importantes de mi formación.

En primer lugar, a mis padres y a mis hermanos, por su apoyo incondicional, su paciencia y su confianza en mí en todo momento. Gracias por estar siempre ahí, incluso en los momentos más complicados, y por darme la motivación necesaria para seguir adelante. También quiero hacer una mención especial a mis amigas del colegio, mis amigas de toda la vida, que me han acompañado en los momentos más intensos de estos últimos años. En especial, a Claudia y a Blanca, por su apoyo constante y por haber estado siempre a mi lado.

A mi ahijada académica, Bea, por su compañía constante y por su capacidad de ver siempre el lado positivo de las cosas. En muchos momentos, su actitud y su forma de afrontar las dificultades han hecho que pareciese que los papeles se invertían, siendo ella quien ejercía el papel de madrina.

A mis tutores, por su orientación, su tiempo y sus valiosas aportaciones a lo largo de todo el desarrollo del trabajo. Asimismo, quiero agradecer a mis mandos y referencias durante estos últimos cinco años, que han contribuido a mi formación de diferentes maneras y han sido una parte fundamental de mi desarrollo.

Por último, a mis compañeros y amigos de la 426-156, por todos los momentos compartidos durante estos años, tanto dentro como fuera del ámbito académico. Gracias por el apoyo, las risas y por hacer que esta etapa sea inolvidable.

A todos vosotros, gracias.

CONTENIDO

Contenido	1
Índice de Figuras	5
Índice de Tablas	11
1 Introducción y objetivos.....	13
1.1 Contexto y motivación.....	13
1.2 Objetivos.....	14
1.3 Estructura de la memoria	14
2 Estado del arte	15
2.1 Datos Sintéticos	15
2.1.1 Necesidad	15
2.1.2 Clasificación y aplicaciones	16
2.2 Técnicas de generación de datos sintéticos	17
2.2.1 Autocodificadores Variacionales (VAE).....	18
2.2.2 Redes Generativas Adversarias (GAN).....	19
2.2.3 <i>Long Short-Term Memory</i> (LSTM).....	21
2.2.4 Grandes Modelos de Lenguaje (LLM)	22
2.3 Técnicas de evaluación de calidad de datos generados	23
2.3.1 Fundamentos principales.....	23
2.3.2 Estadísticas descriptivas	24
2.3.3 Representaciones gráficas	24
2.4 Sistema de Identificación Automática (AIS).....	24
2.4.1 Mensajes AIS	25
2.4.2 Datos AIS estáticos y de viaje.....	26
2.4.3 Datos AIS dinámicos.....	27
2.4.4 Calidad y fiabilidad de los datos AIS	27
2.5 Trabajos relacionados	28
3 Desarrollo del TFG	31
3.1 Recursos materiales	31
3.1.1 Recursos hardware	31
3.1.2 Recursos software.....	32
3.2 Datos de entrenamiento	34
3.2.1 Datos AIS iniciales.....	35
3.2.2 Generación de trayectorias de entrenamiento	35
3.2.3 Datos de entrada	37

3.3 Modelos de generación de datos sintéticos.....	38
3.3.1 Uso del modelo LSTM.....	39
3.3.2 Uso del modelo VAE	43
3.4 Generación de información temporal	51
4 Validación de los resultados	55
4.1 Características generales.....	55
4.1.1 Cargueros.....	56
4.1.2 Petroleros.....	58
4.1.3 Pesqueros.....	60
4.1.4 Pasajeros.....	62
4.1.5 Remolcadores	63
4.2 Comparación de trayectorias generadas por LSTM y datos de trayectorias reales	65
4.2.1 Cargueros.....	66
4.2.2 Pesqueros.....	68
4.3 Comparación de trayectorias generadas con SeqVAE y datos de trayectorias reales	70
4.3.1 Cargueros.....	70
4.3.2 Petroleros.....	72
4.3.3 Pesqueros.....	73
4.3.4 Pasajeros.....	75
4.3.5 Remolcadores	76
4.4 Comparación LSTM y SeqVAE.....	78
4.5 Incorporación de filtrado de trayectorias sintéticas LSTM	78
4.5.1 Cargueros.....	80
4.5.2 Petroleros.....	82
4.5.3 Pesqueros.....	84
4.5.4 Pasajeros.....	86
4.5.5 Remolcadores	88
4.5.6 Inspección visual de las trayectorias LSTM Filtradas.....	90
4.6 Comparación final de los modelos según el tipo de buque	93
5 Conclusiones y líneas futuras.....	95
5.1 Conclusiones.....	95
5.2 Líneas futuras.....	96
6 Bibliografía	97
Anexo I: Implicaciones Sociales, y/o Económicas, y/o Ambientales.....	101
Anexo II: Reflexiones Éticas y Sociales	103

Anexo III: Resultados primera prueba de LSTM para Petroleros, Pasajeros y Remolcadores	105
Anexo IV: Resultados del filtrado de control para LSTM.....	113
Anexo V: Código de generación de trayectorias con LSTM con filtrado de control y filtrado condicionado	125
Anexo VI: Código de generación de trayectorias con SeqVAE	131

ÍNDICE DE FIGURAS

Figura 3-1 Logotipo de Jupyter extraído de [24]	32
Figura 3-2 Diferenciación numérica de los tipos de buques	35
Figura 3-3 Visualización de una sección de los campos de la trayectoria con identificador 1785	37
Figura 3-4 Parámetros iniciales y preprocesado del modelo LSTM	39
Figura 3-5 Código de definición de la función <i>build_sequences</i> del modelo LSTM	40
Figura 3-6 Generación de los datos de entrenamiento del modelo LSTM	41
Figura 3-7 Definición de la arquitectura del modelo LSTM	41
Figura 3-8 Entrenamiento del modelo LSTM	42
Figura 3-9 Variables de las trayectorias generadas	43
Figura 3-10 Función para calcular rumbo	44
Figura 3-11 Definición de la función <i>mover_punto</i>	45
Figura 3-12 Definición de la función <i>reconstruir_trayectorias_sintéticas</i>	45
Figura 3-13 Primera parte de la definición de la arquitectura del modelo SeqVAE	46
Figura 3-14 Segunda parte de la definición del modelo SeqVAE	47
Figura 3-15 Entrenamiento del modelo SeqVAE	47
Figura 3-16 Generación de trayectorias sintéticas con SeqVAE	49
Figura 3-17 Función <i>recuperar_rumbo</i> del modelo SeqVAE	49
Figura 3-18 Recuperación del rumbo y estructuración de variables	50
Figura 3-19 Carga de datos y conversión temporal para la generación de datos temporales	51
Figura 3-20 Cálculo de la distribución de velocidades medias de los datos reales	52
Figura 3-21 Función de cálculo de distancias de las trayectorias sintéticas	53
Figura 3-22 Cálculo de los datos temporales para las trayectorias sintéticas	53
Figura 4-1 Representación de 50 trayectorias reales de Cargueros	57
Figura 4-2 Representación de 50 trayectorias sintéticas generadas con LSTM para Cargueros	57
Figura 4-3 Representación de una selección de trayectorias sintéticas generadas con SeqVAE para Cargueros	58
Figura 4-4 Representación de 50 trayectorias reales de Petroleros	59
Figura 4-5 Representación de 50 trayectorias sintéticas generadas con LSTM para Petroleros	59
Figura 4-6 Representación de una selección de trayectorias sintéticas generadas con SeqVAE para Petroleros	60
Figura 4-7 Representación de 50 trayectorias reales de Pesqueros	61
Figura 4-8 Representación de 50 trayectorias sintéticas generadas con LSTM para Pesqueros	61
Figura 4-9 Representación de una selección de trayectorias sintéticas generadas con SeqVAE para Pesqueros	61

Figura 4-10 Representación de 50 trayectorias reales de Pasajeros.....	62
Figura 4-11 Representación de 50 trayectorias sintéticas generadas con LSTM para Pasajeros	63
Figura 4-12 Representación de una selección de trayectorias sintéticas generadas con SeqVAE para Pasajeros.....	63
Figura 4-13 Representación de 50 trayectorias reales de Remolcadores.....	64
Figura 4-14 Representación de 50 trayectorias sintéticas generadas con LSTM para Remolcadores....	64
Figura 4-15 Representación de una selección de trayectorias sintéticas generadas con SeqVAE para Remolcadores.....	65
Figura 4-16 Distribución velocidades reales Cargueros	66
Figura 4-17 Distribución velocidades LSTM Cargueros	66
Figura 4-18 Distribución tiempo reales Cargueros	66
Figura 4-19 Distribución tiempo LSTM Cargueros.....	66
Figura 4-20 Distribución distancias reales Cargueros	67
Figura 4-21 Distribución distancias LSTM Cargueros	67
Figura 4-22 Distribución geográfica reales Cargueros	67
Figura 4-23 Distribución geográfica LSTM Cargueros.....	67
Figura 4-24 Distribución velocidades reales Pesqueros	68
Figura 4-25 Distribución velocidades LSTM Pesqueros	68
Figura 4-26 Distribución tiempo reales Pesqueros	69
Figura 4-27 Distribución tiempo LSTM Pesqueros.....	69
Figura 4-28 Distribución distancias reales Pesqueros.....	69
Figura 4-29 Distribución distancias LSTM Pesqueros	69
Figura 4-30 Distribución geográfica reales Pesqueros	70
Figura 4-31 Distribución geográfica LSTM Pesqueros	70
Figura 4-32 Distribución velocidades reales Cargueros	71
Figura 4-33 Distribución velocidades SeqVAE Cargueros	71
Figura 4-34 Distribución tiempo reales Cargueros	71
Figura 4-35 Distribución tiempo SeqVAE Cargueros	71
Figura 4-36 Distribución distancias reales Cargueros	71
Figura 4-37 Distribución distancias SeqVAE Cargueros	72
Figura 4-38 Distribución velocidades reales Petroleros	72
Figura 4-39 Distribución velocidades SeqVAE Petroleros.....	72
Figura 4-40 Distribución tiempo reales Petroleros	72
Figura 4-41 Distribución tiempo SeqVAE Petroleros	72
Figura 4-42 Distribución distancias reales Petroleros.....	73
Figura 4-43 Distribución distancias SeqVAE Petroleros.....	73
Figura 4-44 Distribución velocidades reales Pesqueros	74

Figura 4-45 Distribución velocidades SeqVAE Pesqueros.....	74
Figura 4-46 Distribución tiempo reales Pesqueros	74
Figura 4-47 Distribución tiempo SeqVAE Pesqueros	74
Figura 4-48 Distribución distancias reales Pesqueros.....	74
Figura 4-49 Distribución distancias SeqVAE Pesqueros.....	75
Figura 4-50 Distribución velocidades reales Pasajeros.....	75
Figura 4-51 Distribución velocidades SeqVAE Pasajeros.....	75
Figura 4-52 Distribución tiempo reales Pasajeros	75
Figura 4-53 Distribución tiempo SeqVAE Pasajeros	75
Figura 4-54 Distribución distancia reales Pasajeros	76
Figura 4-55 Distribución distancia SeqVAE Pasajeros	76
Figura 4-56 Distribución velocidades reales Remolcadores	77
Figura 4-57 Distribución velocidades SeqVAE Remolcadores	77
Figura 4-58 Distribución tiempo reales Remolcadores.....	77
Figura 4-59 Distribución tiempo SeqVAE Remolcadores.....	77
Figura 4-60 Distribución distancia reales Remolcadores	77
Figura 4-61 Distribución distancias SeqVAE Remolcadores	77
Figura 4-62 Distribución velocidades reales Cargueros	80
Figura 4-63 Distribución velocidades LSTM con filtrado condicionado para Cargueros	80
Figura 4-64 Distribución tiempo reales Cargueros	81
Figura 4-65 Distribución tiempo LSTM con filtrado condicionado para Cargueros.....	81
Figura 4-66 Distribución distancias reales Cargueros	81
Figura 4-67 Distribución distancias LSTM con filtrado condicionado para Cargueros	81
Figura 4-68 Distribución geográfica reales Cargueros	82
Figura 4-69 Distribución geográfica LSTM con filtrado condicionado para Cargueros	82
Figura 4-70 Distribución velocidades reales Petroleros	83
Figura 4-71 Distribución velocidades LSTM con filtrado condicionado para Petroleros	83
Figura 4-72 Distribución tiempo reales Petroleros	83
Figura 4-73 Distribución tiempo LSTM con filtrado condicionado para Petroleros	83
Figura 4-74 Distribución distancias reales Petroleros.....	83
Figura 4-75 Distribución distancias LSTM con filtrado condicionado para Petroleros	83
Figura 4-76 Distribución geográfica reales Petroleros	84
Figura 4-77 Distribución geográfica LSTM con filtrado condicionado para Petroleros	84
Figura 4-78 Distribución velocidades reales Pesqueros	84
Figura 4-79 Distribución velocidades LSTM con filtrado condicionado para Pesqueros	84
Figura 4-80 Distribución tiempo reales Pesqueros	85

Figura 4-81 Distribución tiempo LSTM con filtrado condicionado para Pesqueros	85
Figura 4-82 Distribución distancias reales Pesqueros.....	85
Figura 4-83 Distribución distancias LSTM con filtrado condicionado para Pesqueros	85
Figura 4-84 Distribución geográfica reales Pesqueros	86
Figura 4-85 Distribución geográfica LSTM con filtrado condicionado para Pesqueros	86
Figura 4-86 Distribución velocidades reales Pasajeros.....	86
Figura 4-87 Distribución velocidades LSTM con filtrado condicionado para Pasajeros	86
Figura 4-88 Distribución tiempo reales Pasajeros	87
Figura 4-89 Distribución tiempo LSTM con filtrado condicionado para Pasajeros	87
Figura 4-90 Distribución distancia reales Pasajeros	87
Figura 4-91 Distribución distancia LSTM con filtrado condicionado para Pasajeros.....	87
Figura 4-92 Distribución geográfica reales Pasajeros.....	88
Figura 4-93 Distribución geográfica LSTM con filtrado condicionado para Pasajeros	88
Figura 4-94 Distribución velocidades reales Remolcadores	88
Figura 4-95 Distribución velocidades LSTM con filtrado condicionado para Remolcadores	88
Figura 4-96 Distribución tiempo reales Remolcadores.....	89
Figura 4-97 Distribución tiempo LSTM con filtrado condicionado para Remolcadores	89
Figura 4-98 Distribución distancia reales Remolcadores	89
Figura 4-99 Distribución distancias LSTM con filtrado condicionado para Remolcadores.....	89
Figura 4-100 Distribución geográfica reales Remolcadores	90
Figura 4-101 Distribución geográfica LSTM con filtrado condicionado para Remolcadores	90
Figura 4-102 Representación de 50 trayectorias sintéticas LSTM con filtrado condicionado para Cargueros	91
Figura 4-103 Representación de 50 trayectorias sintéticas LSTM con filtrado condicionado para Petroleros	91
Figura 4-104 Representación de 50 trayectorias sintéticas LSTM con filtrado condicionado para Pesqueros	91
Figura 4-105 Representación de 50 trayectorias sintéticas LSTM con filtrado condicionado para Pasajeros.....	92
Figura 4-106 Representación ampliada de trayectorias sintéticas LSTM con filtrado condicionado para Pasajeros.....	92
Figura 4-107 Representación de 50 trayectorias sintéticas LSTM con filtrado condicionado para Remolcadores.....	92
Figura 4-108 Representación ampliada de trayectorias sintéticas LSTM con filtrado condicionado para Remolcadores.....	93
Figura A3-1 Distribución velocidades reales Petroleros.....	105
Figura A3-2 Distribución velocidades LSTM Petroleros	105
Figura A3-3 Distribución tiempo reales Petroleros	106

Figura A3-4 Distribución tiempo LSTM Petroleros	106
Figura A3-5 Distribución distancias reales Petroleros.....	106
Figura A3-6 Distribución distancias LSTM Petroleros	107
Figura A3-7 Distribución geográfica reales Petroleros.....	107
Figura A3-8 Distribución geográfica LSTM Petroleros	107
Figura A3-9 Distribución velocidades reales Pasajeros.....	108
Figura A3-10 Distribución velocidades LSTM Pasajeros	108
Figura A3-11 Distribución tiempo reales Pasajeros	108
Figura A3-12 Distribución tiempo LSTM Pasajeros	108
Figura A3-13 Distribución distancia reales Pasajeros	108
Figura A3-14 Distribución distancia LSTM Pasajeros	109
Figura A3-15 Distribución geográfica reales Pasajeros.....	109
Figura A3-16 Distribución geográfica LSTM Pasajeros	109
Figura A3-17 Distribución velocidades reales Remolcadores	109
Figura A3-18 Distribución velocidades LSTM Remolcadores.....	109
Figura A3-19 Distribución tiempo reales Remolcadores.....	110
Figura A3-20 Distribución tiempo LSTM Remolcadores	110
Figura A3-21 Distribución distancia reales Remolcadores.....	110
Figura A3-22 Distribución distancias LSTM Remolcadores.....	110
Figura A3-23 Distribución geográfica reales Remolcadores	111
Figura A3-24 Distribución geográfica LSTM Remolcadores.....	111
Figura A4-1 Distribución velocidades reales Cargueros	113
Figura A4-2 Distribución velocidades LSTM con filtrado de control para Cargueros	113
Figura A4-3 Distribución tiempo reales Cargueros	113
Figura A4-4 Distribución tiempo LSTM con filtrado de control para Cargueros	113
Figura A4-5 Distribución distancias reales Cargueros	114
Figura A4-6 Distribución distancias LSTM con filtrado de control para Cargueros.....	114
Figura A4-7 Distribución geográfica reales Cargueros	114
Figura A4-8 Distribución geográfica LSTM con filtrado de control para Cargueros	114
Figura A4-9 Distribución velocidades reales Petroleros.....	115
Figura A4-10 Distribución velocidades LSTM con filtrado de control para Petroleros.....	115
Figura A4-11 Distribución tiempo reales Petroleros	116
Figura A4-12 Distribución tiempo LSTM con filtrado de control para Petroleros	116
Figura A4-13 Distribución distancias reales Petroleros.....	116
Figura A4-14 Distribución distancias LSTM con filtrado de control para Petroleros.....	116
Figura A4-15 Distribución geográfica reales Petroleros.....	116

Figura A4-16 Distribución geográfica LSTM con filtrado de control para Petroleros.....	116
Figura A4-17 Distribución velocidades reales Pesqueros.....	117
Figura A4-18 Distribución velocidades LSTM con filtrado de control para Pesqueros.....	117
Figura A4-19 Distribución tiempo reales Pesqueros	117
Figura A4-20 Distribución tiempo LSTM con filtrado de control para Pesqueros	117
Figura A4-21 Distribución distancias reales Pesqueros.....	118
Figura A4-22 Distribución distancias LSTM con filtrado de control para Pesqueros.....	118
Figura A4-23 Distribución geográfica reales Pesqueros.....	118
Figura A4-24 Distribución geográfica LSTM con filtrado de control para Pesqueros.....	118
Figura A4-25 Distribución velocidades reales Pasajeros.....	119
Figura A4-26 Distribución velocidades LSTM con filtrado de control para Pasajeros.....	119
Figura A4-27 Distribución tiempo reales Pasajeros	119
Figura A4-28 Distribución tiempo LSTM con filtrado de control para Pasajeros.....	119
Figura A4-29 Distribución distancia reales Pasajeros	119
Figura A4-30 Distribución distancia LSTM con filtrado de control para Pasajeros	120
Figura A4-31 Distribución geográfica reales Pasajeros.....	120
Figura A4-32 Distribución geográfica LSTM con filtrado de control para Pasajeros.....	120
Figura A4-33 Distribución velocidades reales Remolcadores	120
Figura A4-34 Distribución velocidades LSTM con filtrado de control para Remolcadores	120
Figura A4-35 Distribución tiempo reales Remolcadores.....	121
Figura A4-36 Distribución tiempo LSTM con filtrado de control para Remolcadores.....	121
Figura A4-37 Distribución distancia reales Remolcadores.....	121
Figura A4-38 Distribución distancias LSTM con filtrado de control para Remolcadores	121
Figura A4-39 Distribución geográfica reales Remolcadores	122
Figura A4-40 Distribución geográfica LSTM con filtrado de control para Remolcadores	122
Figura A4-41 Representación de 50 trayectorias sintéticas LSTM con filtrado de control para Cargueros	123
Figura A4-42 Representación de 50 trayectorias sintéticas LSTM con filtrado de control para Petroleros	123
Figura A4-43 Representación de 50 trayectorias sintéticas LSTM con filtrado de control para Pesqueros	123
Figura A4-44 Representación de 50 trayectorias sintéticas LSTM con filtrado de control para para Pasajeros.....	124
Figura A4-45 Representación de 50 trayectorias sintéticas LSTM con filtrado de control para Remolcadores.....	124

ÍNDICE DE TABLAS

Tabla 2-1 Formato de mensajes normalizados AIS según [17]	25
Tabla 2-2 Nombre y descripción de los datos estáticos	26
Tabla 2-3 Nombre y descripción de los datos de viaje	27
Tabla 2-4 Nombre y descripción de los datos dinámicos	27
Tabla 3-1 Datos estadísticos de las trayectorias finales según el tipo de buque	36
Tabla 3-2 Tiempos de entrenamiento del modelo LSTM según el tipo de buque	42
Tabla 3-3 Tiempos de generación con modelo LSTM según el tipo de buque	43
Tabla 3-4 Tiempos de entrenamiento del modelo SeqVAE según el tipo de buque	48
Tabla 4-1 Estadísticas generales de las trayectorias reales	56
Tabla 4-2 Estadísticas generales de las trayectorias sintéticas LSTM.....	56
Tabla 4-3 Estadísticas generales de las trayectorias sintéticas SeqVAE	56
Tabla 4-4 Velocidades medias según modelo y tipo de buque	79
Tabla 4-5 Distancias medias según modelo y tipo de buque	79
Tabla 4-6 Duración media según modelo y tipo de buque	80
Tabla 4-7 Tiempos de generación con variantes del modelo LSTM según modelo y tipo de buque	80
Tabla 4-8 Modelo de generación óptimo según el tipo de buque	94

1 INTRODUCCIÓN Y OBJETIVOS

1.1 Contexto y motivación

En los últimos años, el análisis del tráfico marítimo a partir de datos del Sistema de Identificación Automático (*Automatic Identification System* (AIS)) ha adquirido una gran relevancia en ámbitos como la seguridad marítima, la gestión del tráfico naval y el desarrollo de sistemas basados en inteligencia artificial. La disponibilidad de grandes volúmenes de datos AIS ha permitido el desarrollo de modelos capaces de analizar patrones de navegación, caracterizar el comportamiento de los buques y extraer información relevante para diferentes aplicaciones en el ámbito marítimo.

En este contexto, el Centro Universitario de la Defensa en la Escuela Naval Militar (CUD-ENM) ha desarrollado en los últimos años una línea de investigación centrada en el uso de técnicas de aprendizaje automático aplicadas a datos AIS. Los primeros trabajos de esta línea tenían como objetivo principal el desarrollo de modelos capaces de identificar el tipo de buque a partir del resto de variables disponibles en los mensajes AIS, abordando así un problema de clasificación basado en datos reales.

A partir de estos trabajos iniciales, la investigación ha ido evolucionando hacia el análisis más profundo de las trayectorias marítimas, la extracción de características relevantes y la mejora del rendimiento de los modelos mediante el tratamiento y enriquecimiento de los datos disponibles. Sin embargo, uno de los principales retos identificados en esta línea es la disponibilidad de conjuntos de datos adecuados para el entrenamiento de modelos de aprendizaje automático, especialmente en términos de calidad, cantidad y representatividad.

Dentro de este contexto, el Trabajo Fin de Grado de Hidalgo Sánchez [1] constituye un paso relevante al abordar la generación de datos sintéticos AIS con el objetivo de mejorar la disponibilidad de datos para el entrenamiento de modelos. En dicho trabajo, la generación se centra en datos estáticos de carácter tabular.

Sin embargo, el análisis del comportamiento de los buques requiere trabajar con datos dinámicos, donde la información se presenta como secuencias temporales que describen la evolución de la posición y otras variables de navegación. Este tipo de datos introduce una complejidad adicional, ya que no solo es necesario modelar relaciones entre variables, sino también las dependencias temporales y la coherencia espacial de las trayectorias.

Como consecuencia, la metodología empleada en trabajos previos basados en datos estáticos no puede aplicarse directamente al problema de generación de trayectorias dinámicas. Esto hace necesario el estudio de nuevas aproximaciones capaces de trabajar con datos secuenciales y generar trayectorias completas que reproduzcan de forma realista el comportamiento de los buques.

En este sentido, a lo largo de este trabajo se han analizado diferentes enfoques y modelos de generación de datos sintéticos, con el objetivo de identificar aquellas metodologías más adecuadas para la generación sintética de trayectorias AIS. Este análisis permite abordar el problema teniendo en cuenta las particularidades de los datos dinámicos y las necesidades específicas del entorno marítimo.

De este modo, el presente Trabajo Fin de Grado se enmarca en la línea de investigación desarrollada en el CUD-ENM, extendiendo los trabajos previos desde el ámbito de la clasificación basada en datos AIS estáticos hacia la generación sintética de trayectorias dinámicas completas, contribuyendo así al desarrollo de modelos más avanzados para el análisis del tráfico marítimo.

1.2 Objetivos

El objetivo principal de este Trabajo Fin de Grado es desarrollar y evaluar modelos capaces de generar datos AIS dinámicos sintéticos a partir de un entrenamiento con trayectorias reales de buques utilizando técnicas de aprendizaje automático.

Para alcanzar este objetivo general se plantean los siguientes objetivos específicos:

- Revisión del estado del arte de las técnicas de generación de datos sintéticos asimilables a datos AIS dinámicos.
- Analizar un conjunto de datos AIS reales con el fin de comprender las características de las trayectorias de navegación y de las variables dinámicas asociadas.
- Realizar el preprocesado y preparación de los datos para su utilización en modelos de aprendizaje automático.
- Implementar modelos generativos basados en aprendizaje profundo capaces de generar trayectorias sintéticas de buques.
- Analizar las trayectorias generadas mediante diferentes métricas relacionadas con variables dinámicas como la velocidad, la distancia recorrida o la duración de las trayectorias.
- Comparar las características de las trayectorias sintéticas con las trayectorias reales para evaluar la capacidad de los modelos para reproducir el comportamiento observado en los datos AIS de entrenamiento.

1.3 Estructura de la memoria

Una vez introducido el contexto y los objetivos del trabajo, a continuación, se describe brevemente la estructura de la memoria.

- **Capítulo 1: Introducción y Objetivos.** Se presenta el contexto general del trabajo, la motivación del estudio y los objetivos que se pretenden alcanzar.
- **Capítulo 2: Estado del arte.** Se describen los conceptos teóricos relacionados con el sistema AIS, el análisis de trayectorias marítimas y las técnicas de aprendizaje automático empleadas en este trabajo.
- **Capítulo 3: Desarrollo del TFG.** Se presenta la metodología empleada para el tratamiento de los datos AIS y la implementación de los modelos utilizados para la generación de trayectorias sintéticas.
- **Capítulo 4: Validación de los resultados.** Se muestran los resultados obtenidos a partir de los modelos implementados y se realiza un análisis comparativo entre las trayectorias reales y las trayectorias generadas.
- **Capítulo 5: Conclusiones y líneas futuras.** Se presentan las principales conclusiones del trabajo y se plantean posibles líneas de trabajo futuro.

Se incluyen a continuación una sección de bibliografía y seis anexos.

2 ESTADO DEL ARTE

En este capítulo se presentan los fundamentos teóricos y técnicos necesarios para el desarrollo del presente Trabajo Fin de Grado.

En primer lugar, se describen los datos sintéticos, su necesidad y principales aplicaciones. A continuación, se revisan las principales técnicas de generación de datos sintéticos y los métodos para evaluar la calidad de los mismos. Seguidamente, se describen las características de los datos AIS, así como su estructura y limitaciones, ya que constituyen la base sobre la que se construye todo el análisis posterior.

Finalmente, se revisan trabajos previos relevantes en el ámbito del análisis de datos AIS, situando el presente trabajo dentro del contexto de investigación actual y justificando la elección de las metodologías empleadas.

Este conjunto de conceptos proporciona el marco teórico necesario para comprender las decisiones metodológicas adoptadas y los resultados obtenidos en los capítulos posteriores.

2.1 Datos Sintéticos

2.1.1 Necesidad

La generación de datos sintéticos surge como respuesta a la necesidad de disponer de grandes volúmenes de datos (*Big Data*) para el desarrollo de trabajos de investigación y entrenamiento de modelos de aprendizaje automático. Por un lado, existe una escasez de datos reales que sean válidos y completos, especialmente en dominios sensibles o altamente especializados. Por otro lado, en la actualidad se observa una dificultad real para acceder a grandes bases de datos, ya que su uso suele estar restringido.

Estas restricciones, generalmente amparadas en criterios éticos, legales y de protección de la privacidad, se deben principalmente a que los datos manejados contienen información de carácter personal, confidencial o sensible. Adicionalmente, otra problemática relevante en la obtención de grandes conjuntos de datos es la fiabilidad y calidad de la información disponible. Datos procedentes de un mismo dominio pueden almacenarse en distintos formatos al ser recogidos desde múltiples fuentes, y con frecuencia presentan errores, valores ausentes o inconsistencias que los hacen parcialmente inutilizables [2].

Esta situación genera la necesidad de llevar a cabo procesos adicionales de recopilación, depuración, procesamiento y validación de los datos, lo que implica un elevado coste computacional, económico y temporal. Como consecuencia, se dificulta la disponibilidad de datos de calidad y se limita la posibilidad

de alcanzar el volumen necesario para entrenar y validar adecuadamente modelos de análisis y aprendizaje automático [2].

Ante la situación actual, como consecuencia de las limitaciones expuestas anteriormente, se ha considerado que la generación de datos sintéticos constituye una de las soluciones más adecuadas para abordar estos problemas. Este enfoque permite disponer de conjuntos de datos artificiales que, aun siendo ficticios, conservan la estructura estadística, las relaciones internas y los patrones relevantes de los datos reales, sin comprometer información sensible ni la privacidad de los individuos. Además, la generación de datos sintéticos permite paliar la escasez de datos mediante técnicas de aumento de datos, disponiendo de esta manera de datos a priori ilimitados.

Al complementar los datos reales con muestras sintéticas que incorporan nuevas variaciones y escenarios se mejora el conjunto inicial, lo que se traduce en un mejor rendimiento en contextos diversos. Asimismo, los datos sintéticos facilitan el intercambio de información y la colaboración entre organizaciones, al ofrecer una alternativa que preserva la confidencialidad, fomenta la transparencia y promueve una mayor apertura en la investigación basada en datos [3].

2.1.2 Clasificación y aplicaciones

Los datos sintéticos pueden presentarse en múltiples formatos en función del tipo de información que se desee representar y del ámbito de aplicación, lo que da lugar a una clasificación basada en su naturaleza. A continuación, se detallarán, tal como se indica en [4], los tipos de datos generados más habituales y relevantes para este trabajo marcando las aplicaciones principales de cada uno.

Entre los más habituales se encuentran los datos sintéticos de tipo audiovisual. Estos datos incluyen audio e imágenes sintéticas. Los **datos de imágenes** representan información visual en forma de píxeles organizados en dos o tres dimensiones, pudiendo incluir canales adicionales como color, profundidad o espectro. Este tipo de datos captura patrones espaciales complejos y es especialmente relevante en aplicaciones donde la percepción visual es clave. Las principales aplicaciones incluyen visión por computador, reconocimiento facial, conducción autónoma, inspección industrial, teledetección y diagnóstico médico por imagen (radiografías, resonancias, TAC) [3]. En el contexto de datos sintéticos, las imágenes generadas artificialmente se utilizan ampliamente para entrenamiento y validación de modelos cuando el etiquetado manual es costoso o existen restricciones de privacidad.

Los **datos de audio** representan señales acústicas en el dominio temporal (y frecuencial tras transformaciones), capturando información relacionada con voz, sonidos ambientales o música. Este tipo de dato es fundamental en tareas de procesamiento del habla y del sonido, donde la variabilidad temporal y tonal es crítica. Sus aplicaciones principales incluyen reconocimiento automático del habla, síntesis de voz, identificación de locutores, detección de eventos acústicos, asistentes virtuales y sistemas de vigilancia sonora. La generación de audio sintético permite ampliar conjuntos de entrenamiento, simular escenarios poco frecuentes y preservar la identidad de los hablantes.

Los **datos de texto** consisten en secuencias de símbolos lingüísticos (palabras, frases o documentos) que representan información semántica y sintáctica. Son altamente estructurados a nivel lógico, aunque no numéricos de forma directa, y su interpretación depende fuertemente del contexto. Las aplicaciones más comunes se encuentran en el procesamiento del lenguaje natural, como clasificación de documentos, análisis de sentimiento, extracción de información, traducción automática, *chatbots* y generación de contenido. El uso de texto sintético es especialmente relevante para entrenar modelos de lenguaje, crear datos balanceados y simular conversaciones o documentos sin exponer información sensible.

Los **datos tabulares** son aquellos que se organizan en filas y columnas, donde cada fila representa una observación y cada columna una variable, pudiendo ser numérica, categórica u ordinal [2]. Es uno de los formatos más comunes en entornos empresariales y científicos, ya que refleja bases de datos relacionales y hojas de cálculo. Sus aplicaciones incluyen análisis estadístico, predicción de riesgo, detección de fraude, análisis financiero, estudios clínicos y toma de decisiones basada en datos. La

generación de datos tabulares sintéticos es especialmente valiosa para compartir información sensible, realizar pruebas de modelos y aumentar conjuntos de datos sin comprometer la privacidad.

Por último, los **datos de trayectorias** representan la evolución espaciotemporal de entidades móviles mediante secuencias ordenadas de posiciones, rumbos y velocidades. Este tipo de dato es fundamental para modelar comportamientos dinámicos y patrones de movilidad. Entre sus principales aplicaciones se encuentran el seguimiento de vehículos y personas, el análisis de tráfico marítimo y aéreo, la movilidad urbana, la robótica y la detección de anomalías en sistemas de vigilancia [5]. La generación de trayectorias sintéticas permite simular escenarios complejos, entrenar algoritmos de predicción y análisis, además de evaluar sistemas en condiciones controladas y reproducibles.

El desarrollo y la aplicación de los datos sintéticos se encuentra estrechamente ligado al campo del **aprendizaje automático o *Machine Learning***. Este término engloba un conjunto de métodos y algoritmos computacionales que permiten a los sistemas informáticos identificar, aprender y modelar patrones y relaciones a partir de datos, sin necesidad de definir explícitamente cada regla de funcionamiento. A partir de los datos disponibles, estos modelos son capaces de realizar inferencias, generar predicciones o producir nuevos datos que preservan las propiedades estadísticas y estructurales del conjunto original [2].

Para que los modelos de aprendizaje automático puedan capturar adecuadamente los patrones presentes en un problema determinado, es necesario disponer de conjuntos de datos amplios, representativos y de alta calidad. La disponibilidad de grandes volúmenes de información permite que los algoritmos identifiquen regularidades, relaciones entre variables y comportamientos complejos que, de otro modo, resultarían difíciles de modelar.

Sin embargo, en numerosos ámbitos la obtención de conjuntos de datos adecuados constituye un desafío significativo. En muchos casos, los datos disponibles son limitados, incompletos o presentan restricciones de acceso debido a motivos legales, operativos o de confidencialidad. Estas limitaciones pueden afectar directamente al rendimiento de los modelos de aprendizaje automático, ya que reducen la capacidad de los algoritmos para aprender representaciones fiables del fenómeno analizado.

En este contexto, la generación de datos sintéticos se presenta como una herramienta especialmente útil. Mediante técnicas de generación artificial de datos es posible ampliar los conjuntos de entrenamiento disponibles, preservando las propiedades estadísticas de los datos reales sin exponer información sensible. De esta forma, los datos sintéticos permiten mejorar el rendimiento de los modelos de aprendizaje automático, facilitar el intercambio de información entre organizaciones y favorecer el desarrollo de sistemas de análisis basados en datos en aquellos entornos donde el acceso a información real se encuentra limitado.

2.2 Técnicas de generación de datos sintéticos

Las técnicas de generación de datos se pueden clasificar de diferentes maneras. En [3] se diferencian dos categorías; métodos estadísticos y modelos de *deep learning* (aprendizaje profundo).

Los **métodos estadísticos para la generación de datos sintéticos** se basan en analizar el comportamiento global de los datos reales con el fin de reproducirlo mediante modelos matemáticos. Estos enfoques parten de la identificación de la distribución que mejor representa los datos originales y del cálculo de sus características estadísticas más relevantes, a partir de las cuales se generan nuevas observaciones artificiales. El procedimiento habitual consiste, primero, en ajustar un modelo estadístico a los datos disponibles y, posteriormente, en extraer muestras a partir de dicho modelo mediante técnicas de muestreo controlado. De este modo, los datos sintéticos obtenidos mantienen propiedades clave de los datos reales, como su distribución de probabilidades, las relaciones entre variables y los principales estadísticos descriptivos, lo que permite que resulten comparables a los datos originales desde un punto de vista estadístico, especialmente en análisis basados en distribuciones agregadas.

La generación de datos sintéticos mediante *Gaussian Noise* se considera un método de generación estadístico y consiste en crear nuevas muestras a partir de datos existentes añadiendo pequeñas perturbaciones aleatorias que siguen una distribución normal o gaussiana. En este enfoque, a cada observación original se le suma un término de ruido generado con media cero y una desviación estándar determinada, lo que produce nuevas instancias que conservan las características estadísticas principales del conjunto de datos inicial, pero introducen una ligera variabilidad. Este método se utiliza habitualmente como técnica de aumento de datos (*data augmentation*), ya que permite ampliar conjuntos de datos reducidos y mejorar la capacidad de generalización de los modelos de aprendizaje automático frente a pequeñas variaciones o ruido presentes en datos reales, aunque su capacidad de generación se limita a producir variaciones locales de los datos originales.

Los **métodos de generación de datos sintéticos basados en modelos de aprendizaje profundo** se apoyan en redes neuronales capaces de aprender representaciones complejas y de alta dimensionalidad a partir de grandes volúmenes de datos reales. Entre estos enfoques destacan los Autocodificadores Variacionales (*Variational AutoEncoder*, VAE), que aprenden una representación latente compacta de los datos y permiten generar nuevas muestras mediante la exploración controlada de dicho espacio, produciendo instancias que conservan las características esenciales de los datos originales. De forma complementaria, las Redes Generativas Adversarias (*Generative Adversarial Networks*, GAN) emplean un esquema de entrenamiento competitivo entre dos redes neuronales, lo que posibilita la generación de datos sintéticos con un alto grado de realismo y diversidad. En el caso de datos secuenciales o series temporales, también se emplean arquitecturas recurrentes como los modelos *Long Short-Term Memory* (LSTM), que permiten capturar dependencias temporales presentes en los datos y generar nuevas secuencias que reproducen las dinámicas observadas en los registros originales. Más recientemente, los grandes modelos de lenguaje (*Large Language Model*, LLM) han ampliado este enfoque al ámbito textual, demostrando una notable capacidad para producir datos sintéticos coherentes y contextualmente consistentes a partir de instrucciones específicas. En conjunto, estos modelos de aprendizaje profundo permiten capturar patrones complejos, relaciones no lineales y estructuras subyacentes difíciles de modelar mediante métodos tradicionales, lo que los convierte en herramientas especialmente adecuadas para la generación de datos sintéticos en dominios como la visión por computador, el procesamiento del lenguaje natural y el aumento de conjuntos de datos para el entrenamiento de modelos avanzados.

2.2.1 Autocodificadores Variacionales (VAE)

Los Autocodificadores (*AutoEncoders*) son un tipo de red neuronal diseñado para el aprendizaje de representaciones de datos de forma no supervisada. Estas arquitecturas constan de dos componentes principales:

- Un codificador, encargado de transformar los datos de entrada en una representación comprimida de menor dimensión captando las características más relevantes de cada dato reagrupándolos juntos según la similitud de sus características en un espacio latente.
- Un decodificador, cuya función es reconstruir los datos originales a partir de dicha representación latente.

En [6] se describen diversas aplicaciones de los **autocodificadores tradicionales (AE)**. Entre las más relevantes se encuentran las siguientes:

- Reducción de la dimensionalidad: Los AE permiten representar datos de alta dimensión en espacios latentes compactos, facilitando tanto su visualización como el posterior procesamiento mediante algoritmos de aprendizaje automático.
- Extracción automática de características: Se debe entrenar al codificador para que sea capaz de detectar y extraer los aspectos esenciales y característicos de los datos. Este proceso permitirá una mejor clasificación y agrupación de los datos similares. En reconocimiento de imágenes esto es especialmente relevante ya que permite detectar patrones visuales importantes.
- Eliminación de ruido: Una vez comprimidos los datos, en el proceso de decodificación, los AE son capaces de extraer de manera limpia y libre de ruido los datos originalmente corruptos. Esta

aplicación es especialmente relevante en el procesamiento de imágenes y señales de audio pudiendo mejorar la calidad de estas.

- Detección de anomalías: Ya que las observaciones que se apartan del comportamiento normal suelen presentar errores de reconstrucción significativamente mayores.

No obstante, los AE tradicionales presentan limitaciones importantes, especialmente en lo relativo a la generación de nuevos datos, ya que su espacio latente carece de una estructura probabilística explícita. Para superar estas limitaciones surgen los *Variational AutoEncoders* (VAE), que introducen un enfoque probabilístico en el proceso de codificación. En lugar de aprender representaciones deterministas, los VAE modelan el espacio latente como una distribución continua, de la cual se extraen muestras durante el proceso de decodificación. Esta formulación permite no solo reconstruir los datos de entrada, sino también generar nuevas instancias sintéticas con características similares a las observadas durante el entrenamiento.

Los VAE mantienen la arquitectura básica de codificador y decodificador de los AE, pero incorporan una formulación probabilística que les permite funcionar como modelos generativos. En lugar de producir una representación latente determinista, el codificador estima los parámetros de una distribución latente (habitualmente una media y una varianza [7]) que aproxima la distribución posterior de las variables latentes condicionada a los datos de entrada. Para posibilitar el entrenamiento mediante retropropagación, los VAE emplean el método de reparametrización, que separa el proceso de muestreo de los parámetros aprendidos por la red.

El modelo generativo lo aporta el decodificador que reconstruye los datos originales o genera nuevas muestras a partir de las variables latentes muestreadas. El entrenamiento del modelo se basa en una función objetivo que combina la calidad de la reconstrucción con un término de regularización que fuerza a la distribución latente aprendida a aproximarse a una distribución previa. Esta regularización impone continuidad y estructura al espacio latente, permitiendo interpolaciones suaves y una generación coherente de nuevas instancias, lo que convierte a los VAE en una herramienta especialmente adecuada para la generación de datos sintéticos y el modelado de patrones complejos.

En [6] se describen también las diferentes aplicaciones de los VAE. Entre las más relevantes se encuentran las siguientes:

- El modelado generativo: La principal ventaja de los VAE es la capacidad de generar nuevas muestras de datos que mantengan los patrones y las características de los originales, aportando ligeras variaciones.
- Detección de anomalías: Los VAE estudian la distribución probabilística de los datos por lo que detectan aquellos que se comportan de manera anómala.
- Imputación de datos y eliminación de ruido: Los VAE permiten, además de la eliminación de ruido como los autocodificadores tradicionales, la reconstrucción de partes ausentes de imágenes tomando muestras de la distribución latente aprendida.

Mientras que los AE tradicionales se centran en la compresión y el aprendizaje de características, los VAE aportan una capacidad adicional al permitir la generación controlada de datos mediante la exploración del espacio latente. Esta propiedad convierte a los VAE en una herramienta especialmente adecuada para la generación de datos sintéticos en dominios complejos, donde resulta necesario capturar patrones no lineales y estructuras subyacentes difíciles de modelar mediante métodos tradicionales.

2.2.2 Redes Generativas Adversarias (GAN)

A pesar de las capacidades generativas que introducen los VAE, estos modelos presentan ciertas limitaciones en cuanto a la fidelidad y nitidez de los datos generados, especialmente en aplicaciones donde se requiere un alto grado de realismo. La formulación probabilística que caracteriza a los VAE, si bien garantiza un espacio latente continuo y estructurado, suele dar lugar a reconstrucciones suavizadas y a una menor capacidad para capturar detalles finos en los datos. Estas limitaciones han motivado el desarrollo de enfoques alternativos dentro del aprendizaje profundo orientados

específicamente a la generación de datos sintéticos. En este contexto surgen las Redes Generativas Adversarias (*Generative Adversarial Networks*, GAN), que abordan el problema de la generación de datos desde una perspectiva diferente.

Las Redes Generativas Adversarias constituyen una de las arquitecturas más influyentes dentro del aprendizaje profundo para la generación de datos sintéticos. Introducidas por Goodfellow en [8], las GAN se fundamentan en un marco de aprendizaje adversarial en el que dos redes neuronales se entrenan simultáneamente con objetivos opuestos. Esta formulación permite modelar distribuciones de datos complejas sin necesidad de definir explícitamente una función de verosimilitud, lo que supone una diferencia clave respecto a otros modelos generativos.

Desde el punto de vista estructural, una GAN está compuesta por dos elementos principales:

- Un generador, el cual recibe como entrada un vector de ruido aleatorio, generalmente muestreado de una distribución simple, y tiene como objetivo producir muestras sintéticas que se asemejen a los datos reales.
- Un discriminador, el cual actúa como un clasificador binario que intenta distinguir entre muestras reales y muestras generadas.

Ambos modelos participan en un juego de suma cero, en el que el generador trata de engañar al discriminador, mientras que este último intenta mejorar su capacidad de detección. El proceso de entrenamiento finaliza cuando el generador produce datos indistinguibles de los reales desde el punto de vista del discriminador.

El funcionamiento de las GAN [8] se basa en la optimización alternada de ambas redes. Durante el entrenamiento, el discriminador se ajusta para maximizar su capacidad de clasificación correcta, mientras que el generador se optimiza para minimizar la probabilidad de que sus muestras sean identificadas como falsas. Este esquema adversarial permite al generador aprender gradualmente la distribución subyacente de los datos reales, dando lugar a muestras sintéticas de alta calidad y realismo.

Las aplicaciones más comunes de las GAN se concentran en dominios donde la fidelidad visual o estructural de los datos es un factor crítico. Entre ellas destacan:

- Generación de imágenes: donde las GAN han demostrado un rendimiento sobresaliente en tareas como la síntesis de rostros, la generación de escenas realistas y la mejora de resolución.
- Aumento de datos: permitiendo enriquecer conjuntos de entrenamiento cuando la disponibilidad de datos reales es limitada o desequilibrada.
- Transferencia de estilo: las GAN permiten transformar imágenes o señales preservando su contenido estructural mientras se modifica su estilo visual o artístico, siendo ampliamente utilizadas en tareas de estilización de imágenes y adaptación de dominios.
- Generación de contenido creativo (arte, música): se emplean para producir obras visuales y composiciones musicales originales, aprendiendo patrones estéticos a partir de grandes colecciones de ejemplos y explorando nuevas combinaciones creativas.
- Síntesis de datos médicos [9]: las GAN se utilizan para generar imágenes médicas y registros clínicos sintéticos que mantienen las propiedades estadísticas de los datos reales, facilitando el entrenamiento de modelos y el intercambio de información sin comprometer la privacidad de los pacientes.
- Generación de datos tabulares y de series temporales [10]: mediante arquitecturas adaptadas, las GAN permiten sintetizar datos estructurados y secuenciales, preservando relaciones entre variables y dependencias temporales, lo que resulta útil en ámbitos como finanzas, movilidad o monitorización de sistemas.

En comparación con otras técnicas de generación de datos sintéticos, las GAN presentan ventajas significativas. En primer lugar, destacan por su capacidad para generar muestras altamente realistas, superando en muchos casos a modelos probabilísticos explícitos como los VAE. Además, al no requerir una formulación explícita de la función de verosimilitud, las GAN pueden adaptarse con mayor

flexibilidad a distribuciones complejas y de alta dimensionalidad. Esta propiedad las convierte en una herramienta especialmente potente para dominios donde la estructura de los datos es difícil de modelar mediante métodos estadísticos tradicionales.

No obstante, las GAN también presentan limitaciones importantes. El proceso de entrenamiento es inherentemente inestable y sensible a la elección de parámetros, pudiendo dar lugar a problemas como que el generador produce un conjunto reducido de muestras con baja diversidad. Además, a diferencia de modelos como los VAE, las GAN no proporcionan una estimación explícita de la probabilidad de las muestras generadas ni un espacio latente estructurado de forma regular, lo que dificulta la interpretación y el control del proceso generativo. En comparación con los métodos estadísticos clásicos, las GAN requieren un mayor coste computacional y grandes volúmenes de datos para alcanzar un rendimiento óptimo.

2.2.3 Long Short-Term Memory (LSTM)

Las redes *Long Short-Term Memory* (LSTM) constituyen una arquitectura específica de redes neuronales recurrentes (RNN) diseñada para el modelado de datos secuenciales y la captura de dependencias temporales de largo alcance. Las RNN convencionales permiten almacenar información de entradas previas mediante conexiones recurrentes, lo que teóricamente posibilita el aprendizaje de patrones temporales. Sin embargo, durante el entrenamiento mediante retropropagación a través del tiempo (*Backpropagation Through Time*, BPTT), los gradientes pueden desaparecer o explotar al propagarse hacia pasos temporales anteriores, lo que dificulta el aprendizaje de relaciones entre eventos muy separados en el tiempo.

El modelo LSTM fue propuesto en [11] para mitigar este problema mediante una arquitectura que permite mantener un flujo de error aproximadamente constante durante la retropropagación. De este modo, la red puede aprender dependencias entre eventos separados por largos intervalos temporales sin perder estabilidad en el proceso de entrenamiento.

La unidad fundamental de una red LSTM es la celda de memoria, cuyo objetivo es almacenar y actualizar información relevante a lo largo del tiempo. Cada celda mantiene un estado interno que representa la memoria de la red en un instante determinado. La actualización de este estado está controlada mediante diferentes puertas, que regulan el flujo de información hacia el interior y exterior de la celda.

En una celda LSTM se distinguen principalmente tres mecanismos de control. La puerta de entrada determina qué información procedente de la entrada actual debe almacenarse en la memoria. La puerta de olvido regula qué parte de la información previamente almacenada debe mantenerse o descartarse. Finalmente, la puerta de salida controla qué información del estado interno se utiliza para producir la salida de la red. Estos mecanismos permiten que el modelo seleccione dinámicamente la información relevante y mantenga memoria de patrones temporales importantes durante largos periodos de tiempo.

La principal ventaja de las redes LSTM frente a otras arquitecturas de aprendizaje automático es su capacidad para modelar dependencias temporales de largo alcance. Gracias a su estructura interna de memoria y al control del flujo de información mediante puertas, estas redes pueden aprender relaciones entre observaciones separadas por numerosos pasos temporales. Otra ventaja relevante es su capacidad para trabajar con datos continuos, representaciones distribuidas y secuencias con ruido, lo que las hace especialmente adecuadas para aplicaciones reales donde las señales presentan variabilidad o incertidumbre.

No obstante, los modelos LSTM también presentan algunas limitaciones. Su arquitectura es más compleja que la de redes neuronales tradicionales, ya que cada celda incorpora múltiples mecanismos de control adicionales, lo que incrementa el número de parámetros del modelo y el coste computacional del entrenamiento. Asimismo, aunque permiten capturar dependencias temporales largas, su rendimiento puede verse limitado en problemas donde la información relevante depende de combinaciones muy específicas de eventos o cuando se requiere una precisión temporal extremadamente elevada.

Debido a su capacidad para mantener memoria de observaciones pasadas, las redes LSTM se utilizan ampliamente para el análisis y la predicción de series temporales. Estas redes pueden aprender patrones dinámicos presentes en secuencias de datos y utilizar la información histórica para estimar la evolución futura de un sistema. Por ejemplo, en aplicaciones relacionadas con el transporte inteligente [12], los modelos LSTM pueden emplearse para predecir trayectorias futuras de vehículos utilizando observaciones previas de su posición, velocidad y del comportamiento de los vehículos circundantes. Este tipo de modelos ha demostrado ser capaz de estimar trayectorias longitudinales y laterales con buena precisión utilizando datos reales de tráfico.

Además de la predicción, los modelos LSTM también pueden utilizarse para generar series temporales sintéticas una vez que han aprendido la dinámica subyacente de los datos. Tras el entrenamiento, el modelo puede producir nuevas secuencias que mantienen propiedades estadísticas y temporales similares a las observadas en los datos originales, lo que resulta útil en tareas de simulación, generación de datos artificiales para entrenamiento y análisis de sistemas dinámicos.

2.2.4 Grandes Modelos de Lenguaje (LLM)

Los Grandes Modelos de Lenguaje (*Large Language Models*, LLMs) son modelos de aprendizaje profundo diseñados para comprender, procesar y generar lenguaje natural en una amplia variedad de tareas. Estos modelos se entrenan sobre grandes volúmenes de datos textuales mediante técnicas de aprendizaje automático no supervisado, lo que les permite capturar patrones lingüísticos complejos y generalizar a tareas no vistas durante el entrenamiento. A diferencia de modelos lingüísticos tradicionales, los LLMs operan con arquitecturas de gran escala y miles de millones de parámetros, lo que explica su elevada capacidad expresiva y su versatilidad en múltiples dominios.

Desde el punto de vista estructural, los LLMs se basan fundamentalmente en la arquitectura *Transformer*, que emplea mecanismos de atención automática para modelar las relaciones contextuales entre los elementos de una secuencia. Este enfoque permite procesar textos largos de forma eficiente, capturando dependencias a corto y largo plazo sin recurrir a estructuras recurrentes. El entrenamiento de los LLMs suele realizarse en dos etapas: una fase de preentrenamiento sobre grandes corpus de texto general y una fase posterior de ajuste fino o adaptación mediante instrucciones o ejemplos específicos, lo que permite especializar el modelo en tareas concretas [13].

Las aplicaciones de los LLMs abarcan un amplio espectro dentro del procesamiento del lenguaje natural, incluyendo traducción automática, generación y resumen de textos, análisis de sentimiento, respuesta a preguntas y sistemas de diálogo. Además, estos modelos se utilizan en la generación de código, asistencia a la toma de decisiones, clasificación documental y extracción de información. Gracias a su capacidad de adaptación, los LLMs han sido incorporados en ámbitos como la educación, el soporte técnico automatizado y el desarrollo de agentes conversacionales avanzados [14].

Entre las principales ventajas de los LLMs destaca su capacidad de generalización, que les permite abordar múltiples tareas sin necesidad de un entrenamiento específico para cada una de ellas. Asimismo, ofrecen una alta calidad en la generación de texto, manteniendo coherencia contextual y fluidez lingüística. Su flexibilidad y escalabilidad los convierten en una herramienta potente para aplicaciones donde se requiere procesamiento avanzado del lenguaje natural.

No obstante, los LLMs presentan limitaciones relevantes. Su entrenamiento y despliegue implican un alto coste computacional y energético, lo que restringe su accesibilidad. Además, pueden generar respuestas incorrectas o inconsistentes, fenómeno conocido como alucinaciones, y reproducir sesgos presentes en los datos de entrenamiento. A ello se suma la baja interpretabilidad de estos modelos, que dificulta comprender el razonamiento interno que conduce a determinadas salidas, planteando desafíos éticos y de fiabilidad en aplicaciones críticas.

En conjunto, los *Large Language Models* representan un avance significativo en el campo de la inteligencia artificial aplicada al lenguaje, combinando arquitecturas basadas en atención con

aprendizaje a gran escala. Sin embargo, su uso efectivo y responsable requiere considerar cuidadosamente sus limitaciones técnicas, computacionales y éticas.

2.3 Técnicas de evaluación de calidad de datos generados

La evaluación de la calidad de los datos sintéticos constituye uno de los aspectos más críticos y, a la vez, más complejos dentro del ámbito de la generación de datos artificiales. A diferencia de los datos reales, cuya validez puede contrastarse directamente con el fenómeno observado, los datos sintéticos requieren mecanismos de evaluación que permitan determinar en qué medida preservan las propiedades estadísticas, estructurales y semánticas del conjunto original, sin comprometer la privacidad. Tal y como se expone en [9], no existe un criterio universal ni una métrica única capaz de evaluar de forma exhaustiva la calidad de los datos generados, ya que dicha calidad depende en gran medida del propósito para el cual se utilizan los datos sintéticos, así como del tipo de datos considerados.

En este contexto, los métodos de evaluación suelen agruparse en diferentes categorías, que abarcan desde comparaciones estadísticas directas entre datos reales y sintéticos, hasta evaluaciones basadas en el rendimiento de modelos de aprendizaje automático entrenados con los datos sintéticos generados. El enfoque adoptado varía según si el objetivo principal es preservar la distribución original de los datos, mantener las relaciones entre variables, mejorar el rendimiento de modelos predictivos o garantizar la privacidad frente a posibles riesgos de divulgación de información sensible.

2.3.1 Fundamentos principales

La evaluación de datos sintéticos se plantea como un problema multidimensional, en el que deben considerarse simultáneamente tres fundamentos principales: fidelidad, diversidad y generalización [9]. Estos criterios no son independientes entre sí y, en muchos casos, presentan compromisos que dificultan su optimización simultánea, lo que refuerza la necesidad de evaluaciones multidimensionales.

La **fidelidad** hace referencia al grado en que los datos sintéticos reproducen las propiedades estadísticas y estructurales de los datos reales originales. Un conjunto sintético con alta fidelidad presenta distribuciones marginales similares, mantiene correlaciones entre variables y conserva patrones relevantes presentes en los datos reales. Este criterio es especialmente importante cuando los datos sintéticos se emplean como sustituto directo de los datos originales, por ejemplo, para análisis estadísticos o entrenamiento de modelos predictivos. No obstante, una fidelidad excesivamente alta puede implicar riesgos desde el punto de vista de la privacidad, ya que el modelo generador podría estar reproduciendo instancias demasiado cercanas a los datos originales.

La **diversidad** evalúa la capacidad del generador para producir una amplia variedad de muestras sintéticas, evitando la repetición de patrones limitados o la generación de datos redundantes. La diversidad es crucial para garantizar que los datos sintéticos cubran adecuadamente el espacio de posibles observaciones y resulten útiles en tareas como el aumento de datos o el entrenamiento de modelos robustos. Sin embargo, un aumento de la diversidad sin control puede llevar a la generación de muestras poco realistas o inconsistentes con el dominio de aplicación.

Por último, la **generalización** se refiere a la capacidad del modelo generador para aprender la distribución subyacente de los datos reales sin memorizar ejemplos concretos. Un buen nivel de generalización implica que los datos sintéticos son novedosos, plausibles y representativos, pero no copias directas de los datos de entrenamiento. Este concepto es clave tanto para garantizar la utilidad de los datos generados como para reducir riesgos relacionados con la divulgación de información sensible. La literatura destaca que la evaluación de la generalización suele abordarse de forma indirecta, mediante métricas de similitud global, pruebas de pertenencia o evaluaciones basadas en el rendimiento de modelos entrenados con datos sintéticos frente a datos reales.

2.3.2 Estadísticas descriptivas

En [9] se destacan las estadísticas descriptivas como el método más simple de evaluación de datos sintéticos. Las estadísticas descriptivas constituyen uno de los enfoques más habituales y accesibles para la evaluación inicial de la calidad de los datos sintéticos, ya que permiten comparar de forma directa propiedades básicas entre los datos reales y los generados. Medidas como la media, mediana, varianza, desviación estándar, valores mínimos y máximos ofrecen una primera aproximación al grado de similitud global entre ambos conjuntos de datos [15]. Cuando estas métricas presentan valores cercanos, puede inferirse que los datos sintéticos reproducen, al menos a nivel agregado, las características fundamentales de la distribución original. No obstante, diversos estudios como [16] advierten que este tipo de análisis resulta insuficiente si se utiliza de forma aislada, puesto que conjuntos de datos con estadísticas descriptivas prácticamente idénticas pueden presentar distribuciones internas muy diferentes y, por tanto, comportamientos dispares en aplicaciones posteriores. En este sentido, las estadísticas descriptivas deben entenderse como una herramienta preliminar, útil para detectar desviaciones evidentes o errores de generación, pero que requiere ser complementada con métodos adicionales para obtener una evaluación más fiable de la calidad de los datos sintéticos.

2.3.3 Representaciones gráficas

Anscombe en [16] hace referencia a las representaciones gráficas como una herramienta fundamental para complementar el análisis estadístico en la evaluación de la calidad de los datos sintéticos, ya que permiten examinar de forma visual la forma y dispersión de las distribuciones generadas. Gráficos como histogramas, diagramas de caja o gráficos de violín facilitan la comparación directa entre datos reales y sintéticos, ayudando a identificar discrepancias en la variabilidad, la presencia de valores atípicos o cambios en la forma de la distribución que no siempre son detectables mediante estadísticas agregadas. Asimismo, los gráficos Q-Q (*Quantile-Quantile plot*) resultan especialmente útiles para evaluar la similitud entre distribuciones, al comparar sus cuantiles de manera sistemática y permitir detectar desviaciones locales o globales respecto al comportamiento esperado. No obstante, en conjuntos de datos de gran dimensión, el análisis visual exhaustivo puede resultar poco escalable, lo que hace necesario combinar estas técnicas con métodos cuantitativos y automatizados que permitan evaluar la calidad de los datos sintéticos de forma más eficiente.

2.4 Sistema de Identificación Automática (AIS)

El Sistema de Identificación Automática (AIS) es un sistema de radiocomunicaciones marítimas normalizado internacionalmente, establecido en el marco del Convenio Internacional para la Seguridad de la Vida Humana en el Mar (SOLAS), cuyo objetivo principal es mejorar la seguridad de la navegación, la vigilancia marítima y el conocimiento del entorno marítimo [5]. El AIS funciona mediante transpondedores instalados a bordo de los buques, los cuales transmiten de forma automática, periódica y en tiempo casi real información relevante del propio buque a otros buques cercanos y a estaciones costeras.

Desde el punto de vista operativo y estratégico, el AIS constituye una fuente masiva de datos (*Big Data*), ya que en áreas de intenso tráfico se generan millones de mensajes diarios. El análisis de estos datos permite construir una imagen casi en tiempo real del tráfico marítimo, facilitando tareas de control, supervisión, prevención de colisiones, gestión del tráfico marítimo y apoyo a operaciones navales y de seguridad. En este contexto, el AIS es un pilar fundamental del concepto de Conciencia del Entorno Marítimo (*Maritime Domain Awareness*), al proporcionar una visión integrada de las actividades que tienen lugar en los espacios marítimos bajo responsabilidad nacional o internacional.

Asimismo, los datos AIS no solo tienen valor inmediato para la vigilancia, sino que también permiten el análisis histórico del comportamiento de los buques, la detección de patrones de navegación, la identificación de rutas habituales y la detección de comportamientos anómalos o potencialmente ilícitos. Por ello, el AIS se ha convertido en un elemento clave para el desarrollo de sistemas avanzados

de análisis basados en técnicas estadísticas y de inteligencia artificial, capaces de procesar grandes volúmenes de información.

2.4.1 Mensajes AIS

Los mensajes AIS contienen tanto información dinámica, como la posición geográfica (latitud y longitud), velocidad sobre el fondo (*Speed Over Ground*, SOG), rumbo o curso sobre el fondo (*Course Over Ground*, COG) y marca temporal, como información estática, que incluye identificadores únicos (*Maritime Mobile Service Identity*, MMSI y número de identificación de la Organización Marítima Internacional, IMO), tipo de buque, dimensiones, estado de navegación y otros datos administrativos. Esta combinación de información permite identificar a cada embarcación y describir su comportamiento dinámico durante la navegación. Además, contienen una serie de datos que proporcionan información sobre el viaje que realiza el buque como la carga, el puerto de salida y el de destino.

Los transpondedores AIS transmiten diferentes tipos de mensajes que permiten a otros buques y a las estaciones en tierra conocer información relevante sobre la plataforma emisora. Aunque la mayoría de estos mensajes están orientados a describir aspectos relacionados con la navegación, el sistema AIS incluye también mensajes destinados a aplicaciones específicas y a la gestión del propio sistema. Según lo definido en [17], los mensajes AIS se clasifican en tres categorías: **mensajes normalizados**, **mensajes específicos de la aplicación (ASM)** y **mensajes de gestión**.

Los **mensajes normalizados** son aquellos encargados de transmitir la información esencial sobre la navegación y las características de las estaciones emisoras. El formato y contenido de cada tipo de mensaje están normalizados por la Unión Internacional de Telecomunicaciones (UIT). En total, se definen 27 tipos de mensajes normalizados, de los cuales resultan especialmente relevantes para este trabajo los recogidos en la Tabla 2-1.

ID del mensaje	Nombre	Descripción
1	Informe de posición	Informe de posición programado
2	Informe de posición	Informe de posición programado y asignado
3	Informe de posición	Informe especial de posición, respuesta a interrogación (equipos móviles a bordo de Clase A)
18	Informe normalizado de posición de equipos de Clase B	Informe de posición normalizado para equipos móviles a bordo de Clase B para que se utilice en vez de los Mensajes 1, 2 o 3
19	Informe de posición ampliado de equipos de Clase B	Informe de posición ampliado para equipos móviles a bordo de Clase B; contiene información estática adicional
27	Mensaje de radiodifusión AIS de largo alcance	Informe programado de posición diseñado para su detección por satélite (implantación a bordo a partir de 2017)

Tabla 2-1 Formato de mensajes normalizados AIS según [17]

Los mensajes 1, 2 y 3 corresponden a estaciones o buques equipados con AIS de Clase A, cuya instalación es obligatoria para determinadas categorías de buques conforme al Convenio SOLAS. Estos equipos se caracterizan por una mayor frecuencia de transmisión, mayor potencia y un nivel de detalle más elevado en la información proporcionada, que incluye datos dinámicos (posición, rumbo y velocidad), datos estáticos (identidad y dimensiones del buque) y datos relativos al viaje (destino, calado y estado de navegación). Este tipo de transpondedores se encuentra habitualmente en buques mercantes de arqueado bruto igual o superior a 300 GT en navegación internacional, buques mercantes de más de 500 GT en navegación nacional, buques de pasaje y grandes buques comerciales.

Por su parte, los mensajes 18 y 19 están asociados a estaciones equipadas con AIS de Clase B, cuya instalación no es obligatoria según SOLAS y que están destinados principalmente a buques de menor tamaño, como embarcaciones de recreo, pequeños pesqueros, buques de trabajo costero y embarcaciones deportivas. Estos equipos presentan características diferentes a las de Clase A, transmitiendo con menor

frecuencia y potencia, y proporcionando un conjunto de información más limitado, en el que generalmente se omiten los datos relativos al viaje.

La segunda categoría de mensajes AIS la constituyen los **mensajes específicos de la aplicación (ASM)**, que permiten la transmisión de información asociada a aplicaciones concretas, como datos meteorológicos, hidrográficos, notificación de cargas peligrosas o requisitos operativos. Estas aplicaciones se identifican mediante un sistema de codificación compuesto por un Código de Área Designado (DAC) y un Identificador de Función (FI), lo que permite su correcta interpretación cuando se dispone del software adecuado. Las administraciones responsables deben registrar estos mensajes ante la IALA, verificar la existencia de aplicaciones similares antes de su implantación y mantener el control sobre su contenido, siendo obligatorio registrar nuevamente cualquier modificación mediante un nuevo DAC y FI.

Por último, los **mensajes de gestión** están destinados al control del comportamiento de las estaciones AIS y a la administración del Enlace de Datos VHF (VDL), un servicio análogo a un sistema de mensajería. Este tipo de mensajes es transmitido exclusivamente por estaciones base, las cuales operan bajo la supervisión de una autoridad competente.

2.4.2 Datos AIS estáticos y de viaje

Los datos AIS estáticos de un buque son aquellos que permanecen invariantes durante la navegación. Este tipo de información describe principalmente las características físicas y estructurales del buque, así como sus datos de identificación, tales como el nombre, el número MMSI, las dimensiones o el tipo de embarcación (Tabla 2-2). Aunque se consideran estáticos, estos datos pueden ser modificados de manera puntual, por ejemplo, en caso de cambio de nombre, bandera o propietario.

Nombre del parámetro	Descripción
Número MMSI	Código de identificación único de 9 dígitos
Número IMO	Código de identificación de 7 dígitos
Nombre y distintivo de llamada	Nombre de hasta 20 caracteres
Eslora y manga	Dimensiones del buque
Tipo de buque	Indicando la función de la embarcación en la mar

Tabla 2-2 Nombre y descripción de los datos estáticos

Por otro lado, los datos AIS de viaje son aquellos que reflejan la intención operativa del buque durante la navegación, como el destino previsto, el calado o el estado de navegación (Tabla 2-3). A diferencia de los datos dinámicos, estos parámetros no se obtienen mediante sensores a bordo, sino que deben ser introducidos manualmente por el operador del sistema AIS. Esta característica la comparten con los datos estáticos.

Nombre del parámetro	Descripción
Calado	La profundidad vertical del buque medida desde la línea de flotación hasta el punto más bajo del casco
Cargamento	Naturaleza de la carga, teniendo especial relevancia las mercancías peligrosas
Puerto de destino/salida	Puerto al que se dirige o desde el cual emprendió la navegación
Estado de navegación	Situación operativa del buque: navegando, fondeado, amarrado, ...
Hora estimada de llegada (ETA)	Fecha y hora previstas de llegadas al destino en UTC (Hora Universal Coordinada)

Tabla 2-3 Nombre y descripción de los datos de viaje

2.4.3 Datos AIS dinámicos

Los datos dinámicos de los buques son aquellos que se caracterizan por su variación continua en el tiempo y por ser generados automáticamente a partir de los sensores a bordo, sin intervención directa del operador. Este conjunto de datos describe el estado cinemático del buque, incluyendo información como la posición, el rumbo y la velocidad, y resulta fundamental para el seguimiento de la navegación y la reconstrucción de trayectorias.

El formato y el contenido de los datos dinámicos transmitidos por las estaciones AIS están **regulados y normalizados**, tal y como se recoge en [17], donde se especifica qué información puede ser emitida por cada tipo de estación. En función de la clase de estación AIS (Clase A o Clase B), los datos transmitidos pueden variar en cuanto a nivel de detalle, frecuencia de actualización y precisión, tal como se ha descrito previamente en el apartado 2.4.1 de este documento. En la Tabla 2-4 se presentan de forma detallada los datos dinámicos más representativos correspondientes a las estaciones de Clase A y Clase B.

Nombre del parámetro	Descripción y Formato
Longitud	Longitud en 1/10.000 min (180°, Este = positivo, Oeste = negativo)
Latitud	Latitud en 1/10.000 min (90°, Norte = positivo, Sur = negativo)
Sello temporal	Según UTC en formato Unix en segundos
SOG	Velocidad sobre el fondo en 1/10 de nudo en un rango de 0 a 102,2 nudos
COG	Rumbo sobre el fondo en 1/10° en un rango de 0 a 359°
Heading o Rumbo verdadero	Grados (0-359)

Tabla 2-4 Nombre y descripción de los datos dinámicos

2.4.4 Calidad y fiabilidad de los datos AIS

El Sistema de Identificación Automática (AIS) constituye una de las principales fuentes de información para el seguimiento y análisis del tráfico marítimo. No obstante, pese a su amplio uso operativo y analítico, los datos AIS no están exentos de errores e inconsistencias que pueden afectar a su fiabilidad y a la calidad de los análisis derivados. Estos errores pueden clasificarse, de forma general, en errores técnicos, errores de transmisión, errores humanos y alteraciones intencionadas [17], cada uno con distintas causas y probabilidades de aparición.

Desde el punto de vista técnico, los sistemas AIS, como cualquier dispositivo radioeléctrico, son susceptibles a interferencias de radiofrecuencia (RFI). Un fenómeno especialmente relevante es el de los trayectos múltiples, que se produce cuando una señal —como la procedente del sistema GPS— llega al receptor a través de varias rutas debido a reflexiones en estructuras metálicas cercanas. Este efecto puede provocar errores significativos en el cálculo de la posición, dando lugar a posiciones desplazadas o incluso a la aparición de blancos fantasmas, es decir, objetivos inexistentes generados por errores de posicionamiento. El AIS, por diseño, no dispone de mecanismos internos para corregir este tipo de errores, por lo que su detección debe realizarse a nivel de sistema o mediante análisis posteriores.

A estas limitaciones se suman los problemas inherentes a la transmisión de datos. El AIS utiliza el enlace de datos VHF, un medio compartido y vulnerable tanto a interferencias involuntarias como a perturbaciones deliberadas. En entornos de alta densidad de tráfico o con mala planificación radioeléctrica, pueden producirse colisiones de mensajes, pérdidas de paquetes y duplicación de información. De hecho, se ha constatado que un mismo mensaje AIS puede ser recibido simultáneamente por múltiples estaciones, lo que genera mensajes duplicados [5] en las bases de datos y requiere procesos de filtrado específicos para su correcta explotación.

Los estudios basados en grandes volúmenes de datos AIS muestran que, aunque la información cinemática —posición, rumbo sobre el fondo (COG) y velocidad sobre el fondo (SOG)— presenta, en general, un alto grado de precisión, otros campos resultan mucho menos fiables. En particular, los datos estáticos y administrativos, como el tipo de buque o el estado de navegación, suelen estar incompletos, mal configurados o directamente ausentes. Se ha observado que una proporción significativa de buques no transmite correctamente su tipo, ya sea por tratarse de estaciones no sujetas al convenio SOLAS (Clase B) o por errores humanos y descuidos operativos incluso en buques de Clase A. Asimismo, se han detectado MMSI duplicados asociados a posiciones geográficas incompatibles, lo que constituye una anomalía grave desde el punto de vista de la integridad de los datos.

Otra fuente relevante de errores proviene de la naturaleza abierta y no segura del AIS. Al tratarse de un sistema de difusión sin mecanismos criptográficos de autenticación, es técnicamente viable la emisión de mensajes falsos o engañosos, ya sea con fines maliciosos o experimentales. Estas suplantaciones pueden generar blancos inexistentes o falsear la identidad y el comportamiento de un buque real. Aunque el AIS permite interrogar a las estaciones para solicitar nuevos reportes, los emisores de mensajes falsos suelen no responder, lo que dificulta su identificación directa. Por ello, se asume que la presencia de reportes AIS falsos es un fenómeno esperable y que debe ser tratado mediante sistemas de monitorización y verificación a nivel de red o de centro de control.

En términos de probabilidad de aparición, los análisis de integridad indican que los errores graves en los datos dinámicos son relativamente poco frecuentes. Estudios basados en modelos estocásticos y procesos de Markov [18] [19] muestran que la integridad de la información dinámica del AIS se sitúa en torno al 99% en condiciones normales, especialmente cuando se utilizan datos GPS o estimaciones por navegación inercial. No obstante, el nivel de fiabilidad puede verse reducido en situaciones que implican maniobras reiteradas, variaciones abruptas del rumbo o una gran separación respecto a las ayudas a la navegación, ya que en estos contextos los modelos de estimación tienden a presentar mayores errores y desviaciones respecto a la trayectoria real.

2.5 Trabajos relacionados

El presente Trabajo de Fin de Grado está encuadrado en una línea de investigación del Centro Universitario de la Defensa en la Escuela Naval Militar (CUD-ENM) orientada al análisis, modelado y explotación de datos marítimos para aplicaciones de vigilancia, navegación y generación de información sintética de alto valor. En este contexto, varios trabajos efectuados en el CUD-ENM han abordado aspectos complementarios que sirven de base y referencia para la presente investigación.

En este contexto, diversos trabajos previos han abordado el estudio de las trayectorias marítimas desde diferentes perspectivas. En primer lugar, el trabajo de Rodríguez Casajús [20] se centra en la

predicción del tipo de buque a partir de datos AIS, utilizando modelos de aprendizaje automático para identificar patrones característicos en las trayectorias. Este enfoque pone de manifiesto la capacidad de los modelos para extraer información relevante a partir de variables dinámicas como la posición, velocidad o rumbo.

En una línea similar, el trabajo de Roldán Calvo [21] profundiza en el uso de representaciones de trayectorias para mejorar la clasificación de buques, introduciendo técnicas más avanzadas de procesamiento de datos y modelado. Este trabajo destaca la importancia de una adecuada representación de las trayectorias para capturar la dinámica del movimiento y mejorar el rendimiento de los modelos predictivos.

Por otro lado, el trabajo de Pavón Jarava [22] aborda la mejora de la eficiencia computacional en sistemas de predicción de tipo de buque, optimizando el uso de parámetros derivados de las trayectorias AIS. Este enfoque introduce una dimensión adicional relacionada con el coste computacional y la eficiencia de los modelos, aspectos clave en aplicaciones operativas donde el volumen de datos es elevado.

Finalmente, el trabajo de Hidalgo Sánchez [1] se centra en la generación de datos sintéticos exclusivamente estáticos para el entrenamiento de modelos supervisados, analizando el impacto de distintas técnicas de generación en el rendimiento de los modelos. Este trabajo pone de relieve la utilidad de los datos sintéticos como herramienta para ampliar conjuntos de datos y mejorar la capacidad de generalización de los modelos de aprendizaje automático.

En este contexto, el presente Trabajo Fin de Grado continúa la línea iniciada en [1] para abordar la generación de información dinámica, más concretamente de la generación trayectorias sintéticas completas a partir de datos AIS reales mediante modelos de aprendizaje profundo. A diferencia de los trabajos anteriores, que utilizan los datos AIS como entrada para tareas de clasificación o predicción, este trabajo aborda el problema inverso: la generación de nuevos datos que reproduzcan las características estadísticas y dinámicas del tráfico marítimo real. De este modo, se contribuye al desarrollo de herramientas que permiten ampliar conjuntos de datos y facilitar el entrenamiento y evaluación de modelos en el ámbito marítimo.

En conjunto, estos trabajos constituyen un corpus de referencia dentro del CUD-ENM que pone de manifiesto la evolución de la investigación en torno al tratamiento, análisis y modelización de datos AIS, proporcionando un marco de apoyo teórico y metodológico para la presente investigación sobre generación y evaluación de datos sintéticos.

Por otro lado, en la literatura reciente sobre generación de datos de trayectorias destaca el trabajo [23] (ST-TrajGAN (*Semantic and Transformer-based Trajectory GAN*)) que presenta un modelo generativo profundo diseñado específicamente para la síntesis de trayectorias preservando la privacidad de los usuarios. Este enfoque surge como respuesta a los riesgos asociados a la publicación y compartición de datos de movilidad, especialmente en servicios basados en localización (*Location-Based Services*, LBS), donde la divulgación directa de trayectorias reales puede revelar ubicaciones sensibles o patrones de comportamiento de los individuos. Con el objetivo de mitigar este problema, ST-TrajGAN propone la generación de trayectorias sintéticas que mantengan las características estadísticas y dinámicas de los datos originales, evitando al mismo tiempo la exposición directa de información personal.

El modelo combina una codificación semántica de las trayectorias con una arquitectura generativa adversarial basada en decodificadores tipo *Transformer*. Cada punto de la trayectoria se transforma en una representación vectorial que incorpora información espacial y contextual, permitiendo al generador aprender patrones espaciotemporales complejos presentes en los datos reales. A partir de estas representaciones, el generador produce trayectorias sintéticas mientras que un discriminador evalúa su plausibilidad frente a trayectorias reales. Además, los autores introducen una función de pérdida específica denominada *TrajLoss*, diseñada para medir la similitud entre trayectorias generadas y reales,

equilibrando la fidelidad de los datos con la preservación de la privacidad. Los resultados experimentales muestran que este enfoque mejora el compromiso entre utilidad analítica y protección de la información sensible, constituyendo una referencia relevante en el ámbito de la generación de datos sintéticos de movilidad y sirviendo como base conceptual para investigaciones posteriores en este campo.

Otro trabajo relevante en el ámbito de la modelización de trayectorias es el propuesto por Altché y De La Fortelle [12], quienes presentan un enfoque basado en redes neuronales recurrentes para la predicción de trayectorias de vehículos en autopistas. En su estudio se plantea el uso de una red *Long Short-Term Memory* (LSTM) para estimar la evolución futura del movimiento de los vehículos a partir de observaciones previas de su posición y velocidad. El objetivo principal de este modelo es mejorar la capacidad de los sistemas de conducción autónoma para anticipar el comportamiento de los vehículos circundantes, permitiendo así tomar decisiones más seguras en situaciones como adelantamientos o cambios de carril.

El enfoque propuesto aprovecha la capacidad de las redes LSTM para modelar series temporales y dependencias a largo plazo, lo que resulta especialmente adecuado para capturar patrones de movimiento en el tráfico. El modelo utiliza información tanto del vehículo objetivo como de los vehículos cercanos, incorporando variables como posiciones relativas, velocidades y tiempo hasta colisión. A partir de estos datos, la red es capaz de predecir la evolución futura de la trayectoria durante varios segundos. Los resultados experimentales muestran que el método alcanza errores medios reducidos en la predicción de la posición lateral y de la velocidad longitudinal, lo que demuestra el potencial de las arquitecturas recurrentes para el análisis y predicción de trayectorias en escenarios de movilidad real.

Este trabajo resulta especialmente relevante para el presente Trabajo de Fin de Grado, ya que demuestra el potencial de los modelos de aprendizaje profundo para analizar y modelar trayectorias a partir de datos de movimiento reales. Aunque el estudio se centra en trayectorias de vehículos en autopistas, el enfoque basado en redes neuronales para capturar patrones espaciotemporales puede trasladarse a otros dominios de movilidad. En particular, la predicción y modelización de trayectorias constituye un problema análogo al abordado en este TFG, donde se analizan trayectorias de buques a partir de datos AIS con el objetivo de comprender su comportamiento y explorar técnicas de generación o modelado de trayectorias sintéticas.

3 DESARROLLO DEL TFG

En este capítulo se describe el desarrollo técnico del Trabajo Fin de Grado, en el que se abordan las diferentes fases necesarias para el tratamiento de datos AIS y la generación de trayectorias sintéticas mediante técnicas de aprendizaje automático.

Para abordar este problema, se han considerado diferentes enfoques orientados al modelado de datos secuenciales, con el objetivo de capturar tanto la evolución temporal como la coherencia espacial del movimiento de los buques. Estos enfoques permiten aprender patrones presentes en los datos reales y generar nuevas trayectorias con características similares.

La estructura del presente capítulo se organiza en varios apartados que describen de forma detallada las distintas fases del desarrollo. En primer lugar, se presentan los recursos materiales utilizados. A continuación, se describen los datos empleados y el proceso seguido para su preparación. Finalmente, se detallan los modelos utilizados y el procedimiento seguido para su entrenamiento y aplicación.

3.1 Recursos materiales

En este apartado se describen los recursos empleados para el desarrollo del Trabajo Fin de Grado, incluyendo tanto el hardware como el entorno software utilizado durante el proceso.

3.1.1 Recursos hardware

El desarrollo del presente Trabajo Fin de Grado se ha llevado a cabo utilizando un equipo de altas prestaciones que permite procesar grandes volúmenes de datos AIS y ejecutar modelos de aprendizaje automático de forma eficiente. Este entorno ha sido empleado tanto en las fases de preprocesado de los datos como en el entrenamiento y evaluación de los modelos.

Las características del recurso hardware utilizado son las siguientes:

- **Procesador:** Intel® Core™ i9-10900K CPU @ 3.70 GHz.
- **Memoria RAM:** 64 GB.
- **Almacenamiento:** 1 TB.
- **Unidad de procesamiento gráfico (GPU):** NVIDIA GeForce RTX 3090 con 24 GB de memoria dedicada.

La disponibilidad de una GPU dedicada permite acelerar significativamente las operaciones necesarias para el entrenamiento de modelos de aprendizaje profundo, especialmente aquellas relacionadas con el cálculo matricial y el procesamiento paralelo de datos.

En cuanto al entorno software, el sistema operativo utilizado ha sido **Ubuntu 22.04 LTS**, proporcionando un entorno estable y adecuado para el desarrollo de aplicaciones de ciencia de datos y

aprendizaje automático. Este entorno facilita la integración de herramientas y bibliotecas especializadas necesarias para el desarrollo del trabajo.

3.1.2 Recursos software

Entorno de trabajo: *Jupyter*

Jupyter es una iniciativa de software libre cuyo objetivo principal es proporcionar herramientas abiertas para la computación interactiva y la ciencia de datos. El proyecto surge como una evolución natural del entorno *Ipython*, desarrollado inicialmente en 2001 por Fernando Pérez como una interfaz interactiva avanzada para el lenguaje Python orientada a la computación científica [24].

En 2014, el proyecto *Ipython Notebook* se escinde formalmente para dar lugar a *Project Jupyter*, con el fin de separar el núcleo interactivo del soporte específico de *Python* y facilitar la expansión hacia otros lenguajes de programación. El nombre “*Jupyter*”, presente en su logotipo (Figura 3-1), hace referencia precisamente a los tres lenguajes que inicialmente se pretendía soportar de manera prioritaria: Julia, *Python* y R.

Jupyter es una iniciativa de software libre orientada al desarrollo de herramientas interactivas para computación científica, ciencia de datos y educación [25]. Su herramienta más conocida, *Jupyter Notebook*, permite combinar código ejecutable, texto explicativo y visualizaciones dentro de un mismo documento interactivo.



Figura 3-1 Logotipo de Jupyter extraído de [24]

Jupyter adopta una arquitectura modular basada en el concepto de kernels, que permiten ejecutar código en distintos lenguajes de programación manteniendo una interfaz web común [24]. Esta arquitectura desacoplada facilita la extensibilidad del sistema y ha permitido que, en la actualidad, *Jupyter* soporte decenas de lenguajes diferentes mediante kernels desarrollados por la comunidad.

La filosofía del proyecto se fundamenta en varios principios clave:

- **Interactividad:** ejecución inmediata del código y visualización de resultados.
- **Reproducibilidad:** posibilidad de documentar y repetir experimentos computacionales.
- **Accesibilidad:** uso mediante navegador web, sin necesidad de entornos complejos.
- **Apertura:** desarrollo abierto y gobernanza comunitaria.

Estos principios han contribuido a la rápida adopción de *Jupyter* tanto en el ámbito académico como en el industrial y ha sido destacado en estudios sobre herramientas de análisis computacional en investigación científica como [25] [26].

En el ámbito de la ingeniería y la ciencia de datos, *Jupyter* facilita el uso integrado de bibliotecas como *NumPy*, *Pandas* o *Matplotlib*, lo que lo convierte en un entorno idóneo para el desarrollo de modelos generativos y análisis de datos.

El uso de *notebooks* permite integrar código, resultados y explicación técnica en un único documento, siguiendo las recomendaciones sobre documentación y reproducibilidad descritas en [27] y alineándose con el concepto de documento computacional presentado en [25].

Bibliotecas de *Python*:

Python es un lenguaje de programación de alto nivel, interpretado y de propósito general, ampliamente empleado en ámbitos científicos, técnicos y de ingeniería. Su sintaxis clara y su ecosistema de bibliotecas especializadas lo convierten en una herramienta especialmente adecuada para el análisis de datos, el desarrollo de modelos de aprendizaje automático y la investigación aplicada. En el contexto de este trabajo, *Python* constituye un elemento central para el procesamiento, análisis y generación de datos AIS sintéticos.

Uno de los principales puntos fuertes de *Python* es su extenso conjunto de bibliotecas orientadas a tareas específicas. Estas bibliotecas permiten abordar de forma eficiente el tratamiento de datos estructurados, el cálculo numérico, la visualización, el análisis geoespacial y la implementación de modelos de aprendizaje profundo. A continuación, se describen las principales herramientas empleadas en el desarrollo del presente trabajo.

- *Pandas* [28]:

Pandas es una biblioteca de *Python* ampliamente utilizada para el análisis y manipulación de datos estructurados. Proporciona estructuras de datos eficientes, principalmente *Series* y *DataFrame*, que permiten trabajar con información tabular y series temporales de forma sencilla y flexible.

Esta biblioteca facilita la lectura, limpieza, transformación y análisis de datos procedentes de distintos formatos, como archivos CSV, Excel o bases de datos. Además, ofrece herramientas específicas para el tratamiento de fechas y tiempos, así como operaciones de filtrado, agregación y combinación de datos.

Gracias a su integración con *NumPy*, *Pandas* permite realizar operaciones vectorizadas de manera eficiente, lo que la convierte en una herramienta fundamental en tareas de preprocesamiento y análisis de datos en proyectos de ciencia de datos y aprendizaje automático.

- *NumPy* [29]:

NumPy proporciona soporte para el cálculo numérico eficiente mediante el uso de arrays multidimensionales y operaciones vectorizadas. Su diseño permite realizar cálculos matemáticos de forma optimizada, reduciendo el coste computacional frente a implementaciones iterativas tradicionales.

Dentro de este proyecto, *NumPy* interviene en el cálculo de distancias, diferencias temporales y operaciones matriciales necesarias para el tratamiento de las trayectorias y la preparación de los datos de entrada a los modelos generativos.

- *Matplotlib* [30]:

Matplotlib es una biblioteca orientada a la representación gráfica de datos en dos dimensiones. Permite generar visualizaciones como gráficos de líneas, diagramas de dispersión e histogramas, facilitando la exploración y análisis de resultados.

En este trabajo se utiliza principalmente para la representación de trayectorias y la comparación visual entre datos reales y sintéticos, contribuyendo al proceso de evaluación de la calidad de los datos generados.

- *Folium* [31]:

Folium está diseñada para la creación de mapas interactivos a partir de datos geoespaciales, apoyándose en la tecnología *Leaflet*. Permite superponer capas, representar puntos y dibujar trayectorias sobre mapas navegables.

Su uso en este proyecto se centra en la visualización espacial de las trayectorias AIS, lo que facilita la interpretación geográfica de los patrones de movimiento y la validación mediante inspección visual de los datos sintéticos generados.

- ***PyTorch*** [32]:

PyTorch es un marco de trabajo para el desarrollo de modelos de aprendizaje profundo basado en el uso de tensores y diferenciación automática. Su estructura dinámica permite definir arquitecturas neuronales complejas de forma modular y adaptable.

En el presente trabajo se emplea para la implementación y entrenamiento de modelos generativos orientados a la síntesis de trayectorias, aprovechando su capacidad de aceleración mediante GPU para optimizar el proceso de aprendizaje.

- ***Tensorflow*** [33]:

TensorFlow es una plataforma para el diseño y entrenamiento de modelos de redes neuronales a gran escala. Basada en grafos computacionales, permite gestionar de manera eficiente operaciones sobre tensores y facilitar el despliegue de modelos en distintos entornos.

Aunque comparte objetivos con *PyTorch*, su arquitectura y enfoque permiten explorar alternativas en el desarrollo de modelos generativos y de aprendizaje automático aplicados al tratamiento de datos de trayectorias.

- ***PyArrow*** [34]:

PyArrow proporciona una interfaz para trabajar con el formato de datos en memoria *Apache Arrow*, optimizado para el procesamiento columnar. Su diseño mejora la eficiencia en la lectura, escritura y transformación de grandes volúmenes de datos.

En este proyecto resulta especialmente útil en las fases de almacenamiento y manipulación de archivos de gran tamaño, contribuyendo a una gestión más eficiente del conjunto de datos AIS.

- ***GeoPandas*** [35]:

GeoPandas amplía las capacidades de *Pandas* para incorporar geometrías espaciales, permitiendo trabajar con puntos, líneas y polígonos junto con sus atributos asociados. Integra operaciones espaciales y gestión de sistemas de referencia geográfica.

Su aplicación en este trabajo está vinculada al tratamiento espacial de las posiciones AIS y a la manipulación de trayectorias en relación con áreas geográficas específicas, asegurando la coherencia cartográfica de los datos.

3.2 Datos de entrenamiento

En esta sección se describen los datos utilizados para el entrenamiento de los modelos desarrollados en este trabajo, así como los procesos de preparación y transformación aplicados sobre ellos. Dado que los datos procedentes de sistemas de monitorización marítima presentan gran volumen, heterogeneidad y posibles inconsistencias, resulta imprescindible realizar un preprocesamiento previo que permita obtener un conjunto de datos estructurado y adecuado para su posterior análisis y utilización en modelos de aprendizaje automático.

Para ello, se parte de los mensajes AIS recopilados, los cuales contienen información dinámica y estática sobre el estado y la posición de los buques. A partir de estos registros se lleva a cabo un proceso de selección, limpieza y transformación de variables con el objetivo de generar trayectorias coherentes y representativas del comportamiento de los buques. Este proceso incluye la organización de los datos, la diferenciación de tipos de buques y la segmentación de trayectorias, dando lugar al conjunto de datos final empleado durante el entrenamiento y evaluación de los modelos.

3.2.1 Datos AIS iniciales

Inicialmente se dispone de un conjunto de datos AIS obtenidos a partir del flujo recibido en tiempo real y durante un mes y medio (entre el 15 de diciembre 2025 y el 30 de enero 2026) del Centro de Operaciones y Vigilancia de Acción Marítima (COVAM). La conexión de los datos se realiza mediante una red privada virtual (VPN) y la información recibida integra dos fuentes diferenciadas: un flujo satélite y un flujo terrestre. El volumen diario de datos asciende aproximadamente a 1,5 GB para el flujo satélite y 800 MB para el flujo terrestre, siendo almacenados en archivos de registro (*logs*) con rotación diaria.

De los distintos tipos de mensajes AIS descritos en el apartado 2.4.1, se procede a la decodificación selectiva de aquellos relevantes para el análisis. En particular, se extraen los mensajes dinámicos tipo 1, 2, 3 y 18, así como los mensajes estáticos tipo 5 y 24. Como resultado del proceso de decodificación, se generan dos archivos en formato CSV diferenciados: uno correspondiente a los datos estáticos y otro a los datos dinámicos. El archivo CSV de datos estáticos contiene los siguientes campos:

msg_type, repeat, mmsi, ais_version, imo, callsign, shipname, ship_type, to_bow, to_stern, to_port, to_starboard, epfd, month, day, hour, minute, draught, destination, dte, timestamp, partno, vendorid model, serial, mothership_mmsi

Por su parte, el archivo CSV de datos dinámicos incluye los campos:

msg_type, repeat, mmsi, status, turn, speed, accuracy, lon, lat, course, heading, second, maneuver, raim, radio, timestamp, reserved_1, reserved_2, cs, display, dsc, band, msg22, assigned

Con el objetivo de reducir el tamaño de los ficheros, se eliminan las columnas consideradas no necesarias para el análisis posterior, con lo que el conjunto de datos dinámicos queda reducido a las siguientes variables:

mmsi, speed, lon, lat, course, heading

Asimismo, se procede a la integración de información relevante de los mensajes estáticos en el conjunto dinámico, incorporando el campo *ship_type* asociado a cada MMSI. Este campo es especialmente relevante ya que, a posteriori, se entrenarán los modelos según los diferentes tipos de buque.

Este preprocesamiento permite trabajar con un conjunto de datos más compacto y enfocado en las variables cinemáticas y de clasificación del buque, facilitando su posterior análisis y utilización en modelos de generación y evaluación de datos sintéticos.

3.2.2 Generación de trayectorias de entrenamiento

El archivo de datos dinámicos CSV se somete a una serie de modificaciones para su mejor explotación y para mejorar la calidad de los datos. En primer lugar, su contenido se vuelca en 5 archivos diferentes identificados por el tipo de buque (siguiendo la referencia mostrada en la Figura 3-2) asignando un identificador numérico diferente a cada uno de los cinco tipos de buque que se van a emplear en el estudio.

```
ship_type_map = {
    "Carguero": 0,
    "Petrolero": 1,
    "Pesquero": 2,
    "Pasajeros": 3,
    "Tug": 4
}
```

Figura 3-2 Diferenciación numérica de los tipos de buques

De esta manera obtenemos los archivos *dyn_0.csv* que contiene los datos dinámicos de todos los cargueros; *dyn_1.csv* con la información de los petroleros; *dyn_2.csv* para pesqueros; *dyn_3.csv* para

pasajeros y, finalmente, *dyn_4.csv* para remolcadores (*tug*). Este proceso de separación es necesario para poder mejorar el entrenamiento y la generación de las trayectorias en función del tipo de buque, como se ha explicado anteriormente.

La generación de trayectorias a partir de los datos dinámicos se realiza de forma individual para cada uno de los archivos generados. En primer lugar, se ordena el *dataframe* cronológicamente en función del campo *timestamp*, garantizando la coherencia temporal de los registros asociados a cada MMSI. Suponiendo la posibilidad de que exista un error en el campo *speed*, a continuación, se calcula la diferencia temporal (campo *dt*) entre mensajes consecutivos y la distancia (campo *distancia*) recorrida entre posiciones sucesivas, lo que permite calcular la velocidad (campo *velocidad*) en nudos a partir de dichas magnitudes.

Con el objetivo de identificar periodos de parada, se define una condición de parada basada en una velocidad inferior a 0,5 nudos mantenida durante un intervalo superior a 10 minutos. Se crea un campo (*state*) indicador de estado, donde 0 representa navegación y 1 indica posible parada. Una parada se considera válida únicamente cuando el criterio temporal mínimo establecido (10 minutos) se cumple.

Posteriormente, se procede a la segmentación de trayectorias, generando una nueva trayectoria cuando se produce alguna de las siguientes condiciones:

- cambio de MMSI,
- cambio de estado (navegación/parada),
- o una separación temporal entre mensajes consecutivos superior a una hora.

Este proceso da lugar a un conjunto de trayectorias clasificadas según su estado dinámico (movimiento o parada). Una vez generadas las trayectorias, se almacenan por número asignado y se aplican una serie de criterios de filtrado para garantizar la coherencia y calidad del conjunto final de datos trabajando únicamente con las trayectorias que cumplen con los siguientes criterios:

- Se establece un mínimo de 100 puntos por trayectoria, descartando aquellas de menor longitud.
- El análisis posterior se restringe exclusivamente a las trayectorias en movimiento.
- Se eliminan aquellas trayectorias cuya velocidad media supere los 40 nudos, al considerarse valores no realistas para los buques analizados.

El resultado es más reducido de trayectorias válidas, cuya distribución se analiza estadísticamente. Se calculan métricas descriptivas (como mediana, media del número de puntos por trayectoria, número máximo y mínimo de puntos) segmentadas según el tipo de buque y cuyos resultados se presentan en la Tabla 3-1. Se consideran trayectorias iniciales a aquellas generadas antes del filtrado y trayectorias finales aquellas que quedan después del filtrado. Todos los datos estadísticos, puntos máximos y mínimos se calculan a partir de las trayectorias finales.

Tipo de buque	Trayectorias iniciales	Trayectorias finales	Media de puntos	Puntos mínimos	Puntos máximos	Mediana de puntos
0-Carguero	606014	7210	1581.69	100	31874	651.0
1-Petrolero	189474	22365	1378.16	100	22019	567.0
2-Pesquero	359359	31163	1187.08	100	35827	390.0
3-Pasajero	99518	12336	698.63	100	23359	315.0
4-Remolcador	150308	14844	755.97	100	23133	341.0

Tabla 3-1 Datos estadísticos de las trayectorias finales según el tipo de buque

Los campos de las trayectorias con las que se trabaja finalmente, dentro de los 5 archivos ya mencionados anteriormente son los siguientes:

lat, lon, timestamp, distancia, dt, velocidad, id_trayectoria

Las trayectorias aparecen entonces divididas en diferentes *dataframes*, como la trayectoria 1785 (mostrada en la Figura 3-3) correspondiente al MMSI 212854000, un buque tipo 4 (remolcador). Esta trayectoria consta de 923 puntos con sus diferencias en distancia, de tiempo y la velocidad media entre puntos.

```

=== Trayectoria 1785 ===
      lat      lon      timestamp      distancia      dt      velocidad
74477  5.454067  5.122055  1.766622e+09  19409.708744  90093.0  0.418783
74478  5.454638  5.122082  1.766622e+09    63.562605    85.0  1.453598
74479  5.456167  5.122063  1.766622e+09   170.030050   257.0  1.286038
74480  5.456910  5.122075  1.766622e+09    82.628508   149.0  1.077966
74481  5.459430  5.122087  1.766623e+09   280.214363   520.0  1.047487
...      ...      ...      ...      ...      ...      ...
75395  5.604717  5.210790  1.766830e+09    3.696139   124.0  0.057941
75396  5.604633  5.210783  1.766832e+09    9.372441  1891.0  0.009634
75397  5.604645  5.210785  1.766832e+09    1.352570   338.0  0.007779
75398  5.604638  5.210785  1.766832e+09    0.778364    32.0  0.047282
75399  5.604633  5.210790  1.766832e+09    0.784390   171.0  0.008917

[923 rows x 6 columns]

```

Figura 3-3 Visualización de una sección de los campos de la trayectoria con identificador 1785

3.2.3 Datos de entrada

Para evaluar adecuadamente la calidad de los datos sintéticos generados por un modelo, es fundamental conocer previamente las características y propiedades del conjunto de datos utilizado durante el entrenamiento. El análisis de los datos de entrada permite identificar aspectos clave como la distribución espacial de las muestras, los rangos de las variables, la densidad de observaciones o la presencia de posibles sesgos en el conjunto original. Estos elementos condicionan directamente el comportamiento del modelo durante el proceso de aprendizaje y, en consecuencia, influyen en la naturaleza de los datos sintéticos que posteriormente se generan.

En este sentido, comprender las características del conjunto de entrenamiento resulta esencial para establecer criterios de evaluación adecuados. La comparación entre datos reales y datos sintéticos no debe realizarse únicamente desde una perspectiva visual o cualitativa, sino también atendiendo a métricas que reflejen si el modelo ha sido capaz de reproducir las propiedades estadísticas fundamentales del *dataset* original. Sin un conocimiento previo de dichas propiedades, resultaría difícil determinar si las trayectorias sintéticas generadas mantienen la coherencia geográfica, las distribuciones espaciales o los patrones presentes en los datos reales.

Además, el análisis preliminar del *dataset* permite detectar posibles limitaciones inherentes a los datos de entrenamiento, como la concentración geográfica de las trayectorias o la ausencia de muestras en determinadas regiones. Estas limitaciones pueden provocar que los modelos aprendan distribuciones incompletas o sesgadas, lo que se reflejará inevitablemente en los resultados sintéticos. Por tanto, identificar estas características desde el inicio facilita una interpretación más rigurosa de los resultados obtenidos y evita atribuir al modelo comportamientos que, en realidad, derivan de las propias restricciones del conjunto de datos.

A partir de este análisis, se procede la definición de una serie de parámetros que sirvan para caracterizar las trayectorias utilizadas en el entrenamiento en función del tipo de buque. Estas métricas permiten comprender mejor el comportamiento de los datos reales y establecer posteriormente una referencia con la que comparar las trayectorias sintéticas generadas por los modelos. Las métricas que se van a analizar son las siguientes:

- En primer lugar, se calcula la **velocidad media de las trayectorias**, lo que permite identificar el comportamiento típico de navegación de cada tipo de buque. Para complementar esta medida, se analiza también la distribución de velocidades, agrupando los valores en varios rangos de velocidad representativos. Este análisis permite observar en qué intervalos se concentra la mayor parte de las observaciones y detectar posibles diferencias entre los distintos tipos de embarcaciones. Como mejora adicional, se incorporan estadísticos descriptivos como la mediana, la desviación estándar y algunos percentiles, lo que permite evaluar la dispersión de las velocidades y la presencia de valores extremos.
- En segundo lugar, se estudia la **distancia recorrida por trayectoria**, calculando tanto la distancia media como la distribución de esta variable. Para ello, las trayectorias se agrupan en distintos rangos de distancia recorrida, lo que facilita analizar la longitud típica de los desplazamientos presentes en el *dataset*.
- De forma complementaria, también resulta relevante analizar la **longitud de las trayectorias en número de puntos**, ya que esta característica está directamente relacionada con la resolución temporal de los datos y puede influir en el comportamiento de los modelos generativos.
- Asimismo, se analiza el **tiempo de duración de las trayectorias**, calculando su valor medio y representando su distribución mediante distintos intervalos temporales. Este análisis permite comprender la escala temporal habitual de los desplazamientos registrados y observar si existen trayectorias significativamente más largas o cortas que la media.
- Además de estas métricas básicas, otra métrica relevante consiste en analizar la **extensión geográfica de las trayectorias**, observando los rangos de latitud y longitud en los que se concentran los datos. Este análisis permite detectar posibles sesgos espaciales en el *dataset* que podrían influir en la capacidad del modelo para generar trayectorias en determinadas regiones.

El análisis conjunto de estas métricas (velocidad, distancia recorrida, duración temporal, longitud de las trayectorias y extensión geográfica) permite obtener una caracterización más completa del comportamiento de los datos reales utilizados durante el entrenamiento. Esta información resulta especialmente valiosa para la evaluación posterior de los datos sintéticos, ya que proporciona múltiples criterios con los que comprobar si las trayectorias generadas por los modelos reproducen de forma coherente las propiedades estadísticas y los patrones de navegación presentes en el *dataset* original.

3.3 Modelos de generación de datos sintéticos

En este apartado se presentan los enfoques utilizados para la generación de trayectorias sintéticas a partir de los conjuntos de datos de entrenamiento. El objetivo es evaluar distintos modelos de aprendizaje profundo capaces de capturar la dinámica temporal del movimiento de los buques y generar nuevas trayectorias que reproduzcan las características observadas en los datos reales.

En una primera fase se consideró la utilización de ST-TrajGAN [23], un modelo basado en Redes Generativas Adversarias diseñado específicamente para la generación de trayectorias espaciotemporales. No obstante, tras analizar su implementación y realizar diversas pruebas preliminares, se decidió descartar este enfoque. La complejidad del código disponible, junto con la dificultad para adaptarlo al conjunto de datos utilizado y la falta de coherencia observada en algunos de los resultados generados, motivaron la búsqueda de alternativas más controlables y explicables dentro del desarrollo del trabajo.

Descartada esta aproximación, se optó por implementar y evaluar dos enfoques basados en modelos secuenciales de aprendizaje profundo, los cuales permiten modelar de forma natural la evolución temporal de las trayectorias marítimas. En primer lugar, se desarrolla un modelo basado en redes LSTM [12], capaz de aprender patrones temporales presentes en las trayectorias. En segundo lugar, se

implementa un modelo SeqVAE (*Sequential Variational Autoencoder*), que permite aprender una representación latente de las trayectorias y generar nuevas secuencias a partir de dicho espacio latente.

3.3.1 Uso del modelo LSTM

En este apartado se emplea un modelo basado en redes neuronales recurrentes de tipo *Long Short-Term Memory* (LSTM) para la generación de trayectorias marítimas. Este tipo de arquitectura, descrita en el apartado 2.2.3, resulta especialmente adecuada para el análisis de series temporales, ya que su estructura de memoria interna permite capturar dependencias a largo plazo entre observaciones consecutivas. El modelo utiliza como entrada una secuencia de observaciones pasadas de la trayectoria del barco, incluyendo variables como latitud, longitud, velocidad, rumbo e intervalo temporal, correspondientes a las últimas posiciones conocidas.

Durante la fase de generación, el modelo se emplea de forma autoregresiva, de modo que cada predicción obtenida se incorpora como nueva entrada para estimar el siguiente punto de la trayectoria. Este procedimiento permite construir la trayectoria futura paso a paso, generando secuencias de posiciones que representan la evolución prevista del movimiento de la embarcación. De esta forma, el modelo no solo realiza estimaciones puntuales, sino que es capaz de generar trayectorias completas a partir de la información histórica disponible.

El bloque de código de la Figura 3-4 define los parámetros iniciales del modelo y realiza el preprocesado de los datos antes de entrenar la red LSTM. En primer lugar, se establecen dos parámetros fundamentales del modelo temporal T_PREV y K_POST . El primero indica el número de observaciones pasadas que se utilizarán como entrada para el modelo, es decir, cada muestra de entrenamiento estará formada por una secuencia de 20 pasos temporales previos. El segundo parámetro define el horizonte de predicción, indicando que el modelo deberá estimar las posiciones de los 5 pasos futuros. Tras realizar diversas pruebas con distintas configuraciones del modelo, se seleccionó finalmente $T_PREV = 20$ para el entrenamiento. Posteriormente, el parámetro K_POST se reajustó durante la fase de generación con el objetivo de que las trayectorias sintéticas tuvieran una longitud equivalente a la media de puntos presente en los datos de entrenamiento. De este modo, si el conjunto de trayectorias de buques pesqueros presenta una media de 1187,08 puntos por trayectoria, las trayectorias sintéticas generadas se producen con 1187 puntos, manteniendo así una estructura comparable a la de los datos reales.

```
T_PREV = 20
K_POST = 5

FEATURE_COLS = ["lat", "lon", "velocidad", "course", "dt"]

df_clean = df_trayectorias_limpias.copy()
df_clean = df_clean[
    ["id_trayectoria", "timestamp", "lat", "lon", "velocidad", "course", "dt"]
]
df_clean = df_clean.dropna()

df_limpio = df_clean.copy()
limpio_stats = {}
EPS = 1e-6
for col in FEATURE_COLS:
    mean = df_limpio[col].mean()
    std = df_limpio[col].std()
    if std < EPS or np.isnan(std):
        std = 1.0
    limpio_stats[col] = (mean, std)
    df_limpio[col] = (df_limpio[col] - mean) / std
```

Figura 3-4 Parámetros iniciales y preprocesado del modelo LSTM

Además, se especifican las variables que formarán parte de la entrada del modelo mediante la lista `FEATURE_COLS`, que incluye las coordenadas geográficas (*lat*, *lon*), la velocidad del barco (*velocidad*), el rumbo (*course*) y el intervalo temporal entre observaciones (*dt*). A partir del conjunto de datos original

(*df_trayectorias_limpias*, cuya obtención se detalla en el punto 3.2.2 de esta memoria), se crea un subconjunto con las columnas relevantes y se eliminan filas en las que no está definido alguno de los valores de las columnas seleccionadas, garantizando así que el modelo trabaje únicamente con datos completos.

Posteriormente se aplica un proceso de normalización de las variables de entrada definidos en *FEATURE_COLS*, con el objetivo de mejorar la estabilidad numérica durante el entrenamiento de la red neuronal. Para cada variable, se calcula la media (*mean*) y desviación estándar (*std*). Estos valores se almacenan en un diccionario *limpio_stats*, lo que permitirá posteriormente desnormalizar las predicciones del modelo y recuperar las coordenadas en su escala original. La normalización se realiza mediante una transformación tipo *z-score*, restando la media y dividiendo por la desviación estándar de cada variable. Además, se introduce un umbral para evitar problemas numéricos en el caso de que alguna variable tenga una desviación estándar extremadamente pequeña o nula; en ese caso, se fija que la desviación tendrá valor 1 para impedir divisiones por cero. Este proceso asegura que todas las variables de entrada tengan escalas comparables, lo cual es especialmente importante para el correcto aprendizaje de redes neuronales recurrentes como las LSTM.

El bloque de código de la Figura 3-5 define la función *build_sequences*, cuyo objetivo es transformar el *dataset* de trayectorias en secuencias temporales adecuadas para entrenar la red LSTM. El modelo no trabaja directamente con filas individuales del *dataset*, sino con ventanas de observaciones ordenadas en el tiempo, por lo que es necesario reorganizar los datos para obtener ejemplos de entrada y salida.

```
#Preparación de datos
def build_sequences(df, feature_cols, t_prev, k_post):
    X, Y_lat, Y_lon = [], [], []

    lat_i = feature_cols.index("lat")
    lon_i = feature_cols.index("lon")

    for _, df_traj in df.groupby("id_trayectoria"):
        df_traj = df_traj.sort_values("timestamp")
        data = df_traj[feature_cols].to_numpy(dtype=np.float32)

        n = len(data)
        if n < t_prev + k_post:
            continue

        for i in range(n - t_prev - k_post):
            X.append(data[i:t_prev+i])
            Y_lat.append(data[t_prev+i:t_prev+i+k_post, lat_i])
            Y_lon.append(data[t_prev+i:t_prev+i+k_post, lon_i])

    return (
        np.array(X, dtype=np.float32),
        np.array(Y_lat, dtype=np.float32),
        np.array(Y_lon, dtype=np.float32)
    )
```

Figura 3-5 Código de definición de la función *build_sequences* del modelo LSTM

Para ello, el código agrupa los datos por *id_trayectoria*, de modo que cada trayectoria se procesa de forma independiente. Posteriormente, cada grupo se ordena por *timestamp* para asegurar que las observaciones siguen el orden temporal correcto. A continuación, se seleccionan las columnas definidas anteriormente en *FEATURE_COLS* y los valores se convierten en un array de *NumPy* para facilitar su uso en el entrenamiento del modelo.

A partir de cada trayectoria se generan ventanas temporales deslizantes. Cada ventana utiliza las posiciones pasadas como entrada del modelo (X), mientras que los pasos siguientes de latitud y longitud se almacenan en Y_{lat} y Y_{lon} , que constituyen la salida que el modelo debe aprender a estimar. Finalmente, todas las secuencias generadas se convierten en arrays de *NumPy*, obteniendo así las estructuras X , Y_{lat} y Y_{lon} que se utilizarán posteriormente para entrenar la red LSTM.

```
#Uso del dataset real
X_train, Y_lat_train, Y_lon_train = build_sequences(
    df_limpio,
    FEATURE_COLS,
    T_PREV,
    K_POST
)
```

Figura 3-6 Generación de los datos de entrenamiento del modelo LSTM

El último paso antes del entrenamiento es usar la función *build_sequences* para generar los datos que se emplearán en el entrenamiento. La función se aplica sobre el *dataframe* *df_limpio*, que contiene las trayectorias tras el proceso de limpieza y normalización de los datos (Figura 3-6).

Como resultado, se obtienen tres arrays: X_{train} , Y_{lat_train} y Y_{lon_train} . El primero contiene las secuencias de entrada del modelo, formadas por tramos de trayectorias pasadas. Los otros dos arrays contienen las coordenadas de latitud y longitud futuras que el modelo debe aprender a estimar. De esta forma, el *dataset* queda preparado en el formato adecuado para entrenar el modelo LSTM.

```
#Modelo LSTM
inputs = layers.Input(shape=(T_PREV, len(FEATURE_COLS)))

x = layers.LSTM(32)(inputs)
x = layers.Dense(16, activation="relu")(x)

lat_out = layers.Dense(K_POST)(x)
lon_out = layers.Dense(K_POST)(x)

model = models.Model(inputs, [lat_out, lon_out])

optimizer = tf.keras.optimizers.Adam(
    learning_rate=1e-4,
    clipnorm=1.0
)

model.compile(
    optimizer=optimizer,
    loss="mse"
)

model.summary()
```

Figura 3-7 Definición de la arquitectura del modelo LSTM

El bloque mostrado en la Figura 3-7 define la arquitectura del modelo LSTM utilizada para el aprendizaje de las trayectorias. La red recibe como entrada secuencias de datos y las procesa mediante una capa LSTM con 32 unidades, que captura las dependencias temporales. A continuación, una capa *Dense* con activación *ReLU* extrae características intermedias, y dos capas de salida generan las predicciones de latitud y longitud futuras.

El modelo se compila utilizando el optimizador *Adam* y la función de pérdida MSE (error cuadrático medio), que mide la diferencia entre las posiciones reales y las estimadas durante el entrenamiento.

El bloque de código de la Figura 3-8 ejecuta el entrenamiento del modelo mediante la función *model.fit*. La red aprende a partir de las secuencias de entrada X_{train} y de las coordenadas reales de salida (Y_{lat_train} y Y_{lon_train}). El entrenamiento se realiza durante un número determinado de

iteraciones (1 *epochs* en el caso de la Figura 3-8), procesando los datos en lotes (*batch_size*) de 64 trayectorias y mezclando las muestras para mejorar la generalización del modelo.

En cuanto al proceso de entrenamiento, tras realizar diferentes pruebas con distintas configuraciones, se observa que los resultados obtenidos presentan una mayor coherencia en la representación visual de las trayectorias al aumentar el número de iteraciones (*epochs*). Cada iteración corresponde a una iteración completa sobre el conjunto de datos de entrenamiento, cuyo tiempo de ejecución se recoge en la Tabla 3-2. Durante las primeras iteraciones se aprecia una mejora progresiva en el comportamiento del modelo, reflejada en una mayor coherencia en las trayectorias generadas y en una reducción del error de predicción, observándose resultados más consistentes hasta aproximadamente 15 iteraciones. A partir de este punto, el aumento del número de iteraciones no produce mejoras significativas en la calidad de las trayectorias generadas, por lo que el incremento en el tiempo de entrenamiento no resulta justificable.

```
#Entrenamiento
model.fit(
    X_train,
    [Y_lat_train, Y_lon_train],
    epochs=1,
    batch_size=64,
    shuffle=True
)
```

Figura 3-8 Entrenamiento del modelo LSTM

Por otro lado, en los conjuntos de datos correspondientes a petroleros (1) y pesqueros (2), el elevado número de puntos que compone cada trayectoria, junto con el gran volumen de trayectorias disponibles, provoca un incremento considerable en los tiempos de entrenamiento. Para mantener un proceso de entrenamiento manejable desde el punto de vista computacional, se decidió reducir a una cuarta parte el número de trayectorias utilizadas como entrada al modelo. De este modo se consigue disminuir el coste computacional del entrenamiento sin comprometer de forma significativa la representatividad del conjunto de datos empleado. Se puede comprobar la diferencia en el número de trayectorias para dichos tipos de buques comparando la Tabla 3-1 y la Tabla 3-2.

Tipo de buque	Trayectorias de entrenamiento	Media de puntos (<i>n</i>)	Tiempo de entrenamiento 1 epoch	Tiempo de entrenamiento 15 epoch
0-Carguero	7210	1581.69	315s	1h 18m 45s
1-Petrolero	5591	1378.16	362s	1h 30m 30s
2-Pesquero	7790	1187.08	427s	1h 46m 45s
3-Pasajero	12336	698.63	381s	1h 35m 15s
4-Remolcador	14844	755.97	376s	1h 34m 00s

Tabla 3-2 Tiempos de entrenamiento del modelo LSTM según el tipo de buque

Finalmente, una vez entrenado el modelo, se guarda dentro del sistema como *modelo_trayectorias_tipodebuque.keras* y se procede a la generación de las trayectorias. De esta manera no es necesario volver a entrenar el modelo para volver a utilizarse. El procedimiento de generación de trayectorias comienza llamando al modelo con la función:

```
model = load_model("modelo_trayectorias_tipodebuque.keras")
```

Después, se selecciona una secuencia inicial de datos históricos que sirve como punto de partida. A partir de esa información, el modelo estima la siguiente posición geográfica del barco. La predicción obtenida se transforma de nuevo a la escala original de las coordenadas para representar posiciones reales.

Finalmente, una vez entrenado el modelo se procede a la generación de las trayectorias. El procedimiento comienza seleccionando una secuencia inicial de datos históricos que sirve como punto

de partida. A partir de esa información, el modelo estima la siguiente posición geográfica del barco. La predicción obtenida se transforma de nuevo a la escala original de las coordenadas para representar posiciones reales.

El proceso se realiza de forma iterativa y autorregresiva. En cada paso, la nueva posición estimada se incorpora a la secuencia temporal que sirve de entrada al modelo, desplazando los datos más antiguos. De esta manera, cada predicción se utiliza como información para generar la siguiente, permitiendo construir progresivamente una trayectoria completa.

Durante el proceso de generación se incorpora una restricción geográfica adicional destinada a garantizar la coherencia geográfica de las trayectorias generadas. Cada punto producido por el modelo es evaluado mediante una comprobación espacial que determina si la posición corresponde a una zona marítima. En el caso de que algún punto de la trayectoria se sitúe sobre tierra firme, la trayectoria generada es descartada. Este proceso de validación espacial permite asegurar que las trayectorias sintéticas obtenidas mediante el modelo LSTM mantienen coherencia geográfica con el comportamiento real de los buques.

Se generan trayectorias sintéticas siguiendo dos procedimientos. En primer lugar, se realiza una prueba inicial del modelo generando 200 trayectorias compuestas por 100 puntos cada una, con el objetivo de verificar el correcto funcionamiento del proceso de generación y la coherencia de los resultados obtenidos. Posteriormente, y una vez validado el comportamiento del modelo, se procede a la generación del conjunto definitivo de trayectorias sintéticas. En esta segunda fase se generan 35 trayectorias con una longitud igual a n , donde n corresponde al número medio de puntos de las trayectorias presentes en los datos de entrenamiento para cada tipo de buque. El tiempo necesario para la generación de estas trayectorias se recoge en la Tabla 3-3, donde se muestran los tiempos de ejecución asociados al proceso de generación.

Tipo de buque	Trayectorias de entrenamiento	Media de puntos (n)	Tiempo de generación 200 tray/100puntos	Tiempo de generación 35 tray/ n puntos
0-Carguero	7210	1581.69	43m 18s	-
1-Petrolero	5591	1378.16	16m 54.13s	1h 18m 8s
2-Pesquero	7790	1187.08	32m 44.61s	-
3-Pasajero	12336	698.63	29m 52,8s	-
4-Remolcador	14844	755.97	1h 11m 47.6s	-

Tabla 3-3 Tiempos de generación con modelo LSTM según el tipo de buque

Finalmente, todos los puntos generados se almacenan en un propio dataframe (*df_todas_trajs*, Figura 3-9) que recoge la identificación de la trayectoria, el paso temporal correspondiente y las coordenadas geográficas estimadas. Esto permite analizar, visualizar o representar posteriormente las trayectorias generadas por el modelo.

```
df_todas_trajs.columns
```

```
Index(['traj_id', 'step', 'lat_pred', 'lon_pred'], dtype='str')
```

Figura 3-9 Variables de las trayectorias generadas

3.3.2 Uso del modelo VAE

En este apartado se presenta un enfoque basado en VAE (mencionado en el apartado 2.2.1 del estado del arte) para la generación de trayectorias sintéticas. A diferencia del modelo LSTM descrito en el apartado anterior, en este caso el entrenamiento no se realiza directamente sobre coordenadas geográficas absolutas, sino sobre variables dinámicas que describen el movimiento del buque entre

posiciones consecutivas. Este enfoque permite representar la trayectoria de forma más estable desde el punto de vista numérico y facilita el aprendizaje de patrones de movimiento por parte del modelo.

Para ello, a partir de las posiciones geográficas consecutivas de cada trayectoria se calcula, en primer lugar, el rumbo de desplazamiento entre dos puntos. El rumbo representa la dirección del movimiento del buque sobre la superficie terrestre y se obtiene a partir de las coordenadas geográficas de dos posiciones consecutivas. La función *calcular_rumbo* (Figura 3-10) recibe como entrada las latitudes y longitudes de ambos puntos (*lat1*, *lon1*, *lat2*, *lon2*). Inicialmente las coordenadas se transforman a radianes, ya que las operaciones trigonométricas utilizadas requieren esta unidad de medida.

```
def calcular_rumbo(lat1, lon1, lat2, lon2):
    lat1 = np.radians(lat1)
    lat2 = np.radians(lat2)
    dlon = np.radians(lon2 - lon1)

    x = np.sin(dlon) * np.cos(lat2)
    y = (np.cos(lat1)*np.sin(lat2) -
        np.sin(lat1)*np.cos(lat2)*np.cos(dlon))

    bearing = np.degrees(np.arctan2(x, y))
    return(bearing + 360) % 360
```

Figura 3-10 Función para calcular rumbo

A continuación, se calcula la diferencia de longitud entre ambos puntos y se emplean relaciones trigonométricas para obtener las componentes intermedias *x* e *y*, que describen la dirección del desplazamiento. El rumbo se determina finalmente mediante la función *arctan2*, que permite calcular el ángulo teniendo en cuenta el cuadrante correcto. El resultado se transforma posteriormente a grados y se normaliza al intervalo $[0^\circ, 360^\circ]$, obteniendo así el rumbo entre ambas posiciones de la trayectoria.

Una vez obtenido el rumbo, es posible estimar la nueva posición geográfica de un punto a partir de una posición inicial, una dirección de desplazamiento y una distancia recorrida. Este proceso se implementa mediante la función *mover_punto* del bloque de código de la Figura 3-11, que calcula una nueva posición geográfica (*lat* y *lon*) considerando la Tierra como una esfera de radio aproximado de 6.371 km.

Esta función permite describir el movimiento de una trayectoria a partir de un punto inicial y estas dos variables (rumbo y velocidad). La función *reconstruir_trayectoria_sintetica* (Figura 3-12), tiene como objetivo es reconstruir una trayectoria geográfica completa a partir de una posición inicial y una secuencia de variables de movimiento, utilizando la función *mover_punto* explicada anteriormente.

La función recibe como entrada una trayectoria *traj*, que contiene para cada paso temporal el seno y coseno del rumbo, la velocidad del buque (*vel*) y el intervalo temporal entre observaciones (*dt*), junto con una posición inicial definida por las coordenadas *lat0* y *lon0*. Inicialmente se establecen estas coordenadas como punto de partida y se crean dos listas destinadas a almacenar las posiciones generadas durante el proceso.

Posteriormente, se recorre la secuencia de variables de movimiento mediante un bucle. En cada iteración se reconstruye el rumbo a partir de su representación trigonométrica mediante la función *arctan2*. Una vez obtenido el rumbo, se calcula la distancia recorrida multiplicando la velocidad por el intervalo temporal correspondiente. Con estos valores se emplea la función *mover_punto* para estimar la siguiente posición geográfica de la trayectoria.

Las nuevas coordenadas obtenidas se almacenan en las listas correspondientes y el proceso se repite para cada paso temporal de la secuencia. Finalmente, las posiciones calculadas se convierten en arrays de *NumPy*, obteniendo así la trayectoria completa reconstruida a partir de las variables dinámicas de movimiento.

```
def mover_punto(lat, lon, rumbo, distancia_m):
    R = 6371000 # radio tierra en metros

    lat = np.radians(lat)
    lon = np.radians(lon)
    rumbo = np.radians(rumbo)

    lat2 = np.arcsin(
        np.sin(lat)*np.cos(distancia_m/R) +
        np.cos(lat)*np.sin(distancia_m/R)*np.cos(rumbo)
    )

    lon2 = lon + np.arctan2(
        np.sin(rumbo)*np.sin(distancia_m/R)*np.cos(lat),
        np.cos(distancia_m/R) - np.sin(lat)*np.sin(lat2)
    )

    return np.degrees(lat2), np.degrees(lon2)
```

Figura 3-11 Definición de la función *mover_punto*

A continuación, se implementa la clase SeqVAE, un modelo basado en Autocodificadores Variacionales (VAE) diseñado para trabajar con datos secuenciales. La arquitectura combina redes recurrentes GRU (*Gated Recurrent Unit*) con un espacio latente probabilístico, permitiendo aprender una representación compacta de trayectorias y posteriormente reconstruirlas o generar nuevas secuencias sintéticas.

```
# Reconstruye una trayectoria de Lat,Lon a partir de trayectoria con Lat,Lon iniciales, y rumbos,
# velocidades y tiempos en cada punto.
def reconstruir_trayectoria_sintetica(traj, lat0, lon0):
    lat = lat0
    lon = lon0

    lats = []
    lons = []

    for rumbo_sin, rumbo_cos, vel, dt in traj:

        rumbo = np.degrees(np.arctan2(rumbo_sin, rumbo_cos))
        rumbo = (rumbo + 360) % 360

        distancia = vel * dt
        #utiliza funcion para siguiente Lat,Lon
        lat, lon = mover_punto(lat, lon, rumbo, distancia)

        lats.append(lat)
        lons.append(lon)

    return np.array(lats), np.array(lons)
```

Figura 3-12 Definición de la función *reconstruir_trayectorias_sintéticas*

Antes de proceder a la definición de la arquitectura del modelo SeqVAE, es necesario preparar los datos que se utilizarán como entrada durante el entrenamiento. A diferencia del modelo LSTM descrito anteriormente, en este enfoque las trayectorias no se representan directamente mediante coordenadas geográficas absolutas, como ya se ha comentado anteriormente, sino mediante variables dinámicas derivadas del movimiento del buque.

Para ello, a partir de las posiciones GPS consecutivas de cada trayectoria se calcula el rumbo entre puntos sucesivos y se generan nuevas variables que describen la dinámica del movimiento, como el seno y el coseno del rumbo, la velocidad y la diferencia temporal entre observaciones. Esta transformación permite eliminar la discontinuidad angular asociada al rumbo y facilita el aprendizaje del modelo.

Posteriormente, las variables se normalizan para garantizar una escala homogénea entre ellas y se organizan en secuencias temporales de longitud fija. Estas secuencias constituyen el conjunto de datos de entrenamiento que se utilizará para alimentar el modelo generativo basado en VAE. Finalmente, las secuencias se convierten en tensores y se agrupan en lotes mediante un *DataLoader*, lo que permite entrenar el modelo de forma eficiente durante el proceso de aprendizaje.

```
class SeqVAE(nn.Module):

    # 4 dimensiones (features), 64 tamaño de los estados ocultos de la GRU, 16 espacio latente del VAE
    def __init__(self, input_dim=4, hidden_dim=64, latent_dim=16):
        super().__init__()

        # Encoder: utilizamos GRU
        self.encoder_rnn = nn.GRU(input_dim, hidden_dim, batch_first=True)
        self.fc_mu = nn.Linear(hidden_dim, latent_dim)
        self.fc_logvar = nn.Linear(hidden_dim, latent_dim)

        # Decoder
        self.decoder_input = nn.Linear(latent_dim, hidden_dim)
        self.decoder_rnn = nn.GRU(hidden_dim, hidden_dim, batch_first=True)

        # capa de salida
        self.output_layer = nn.Linear(hidden_dim, input_dim)

    def encode(self, x):
        _, h = self.encoder_rnn(x)
        h = h[-1] # Tomamos el último estado oculto
        mu = self.fc_mu(h)
        logvar = self.fc_logvar(h)
        return mu, logvar
```

Figura 3-13 Primera parte de la definición de la arquitectura del modelo SeqVAE

En la inicialización del modelo (Figura 3-13) se definen los principales componentes de la arquitectura. El codificador utiliza una GRU para procesar la secuencia de entrada y obtener un estado oculto que resume la información temporal. A partir de este estado se calculan la media (*mu*) y el logaritmo de la varianza (*logvar*), que definen la distribución del espacio latente. Mediante la reparametrización, se genera un vector latente *z* que permite entrenar el modelo mediante retroprogramación.

El decodificador (Figura 3-14) transforma el vector latente en un estado inicial y, mediante otra GRU, genera la secuencia reconstruida. Finalmente, se aplican restricciones a la salida: la tangente hiperbólica limita los valores del seno y coseno del rumbo al rango $[-1,1]$, mientras que la función *softplus* garantiza que la velocidad y el tiempo sean positivos. El método *forward* conecta todo el proceso de codificación, muestreo latente y decodificación para reconstruir la secuencia de entrada.

```
def reparameterize(self, mu, logvar):
    std = torch.exp(0.5 * logvar)
    eps = torch.randn_like(std)
    return mu + eps * std

def decode(self, z, seq_len):
    h0 = self.decoder_input(z).unsqueeze(0)
    # Inicializamos entrada del decoder
    decoder_input = torch.zeros(z.size(0), seq_len, h0.size(-1)).to(z.device)
    out, _ = self.decoder_rnn(decoder_input, h0)

    raw_output = self.output_layer(out)

    #restricciones físicas
    sin_cos = torch.tanh(raw_output[:, :, 0:2])
    vel_time = F.softplus(raw_output[:, :, 2:4])
    return torch.cat([sin_cos, vel_time], dim=-1)

def forward(self, x):
    mu, logvar = self.encode(x)
    z = self.reparameterize(mu, logvar)
    x_recon = self.decode(z, x.size(1))
    return x_recon, mu, logvar
```

Figura 3-14 Segunda parte de la definición del modelo SeqVAE

Una vez definida la arquitectura del modelo SeqVAE, se procede a la configuración del proceso de entrenamiento, visible en el bloque de código de la Figura 3-15. Para ello, se inicializa el modelo y se asigna al dispositivo de cálculo, lo que permite aprovechar aceleración por GPU en caso de estar disponible. A continuación, se define el optimizador, basado en el algoritmo *Adam*, encargado de actualizar los parámetros de la red neuronal durante el entrenamiento con una tasa de aprendizaje (*lr*) igual a 0.001.

```
model = SeqVAE().to(device)
optimizer = torch.optim.Adam(model.parameters(), lr=0.001)

def vae_loss(x_recon, x, mu, logvar):
    recon_loss = nn.MSELoss()(x_recon, x)
    kl = -0.5 * torch.mean(1 + logvar - mu.pow(2) - logvar.exp())
    return recon_loss + 0.001 * kl

epochs = 100

start_time = time.time() # inicio del entrenamiento

for epoch in range(epochs):
    total_loss = 0
    for (X,) in dataloader:
        X = X.to(device)

        x_recon, mu, logvar = model(X)
        loss = vae_loss(x_recon, X, mu, logvar)

        optimizer.zero_grad()
        loss.backward()
        optimizer.step()

    total_loss += loss.item()
```

Figura 3-15 Entrenamiento del modelo SeqVAE

Posteriormente se define la función de pérdida *vae_loss*, que determina cómo se evalúa el rendimiento del modelo durante el entrenamiento. Esta función combina dos componentes principales. El primero es el error de reconstrucción *recon_loss*, calculado mediante el error cuadrático medio entre la secuencia original x y la secuencia reconstruida por el modelo x_{recon} . El segundo término corresponde a la divergencia de *Kullback–Leibler* kl , que actúa como un mecanismo de regularización para que la distribución del espacio latente aprendida por el modelo se aproxime a una distribución normal.

A continuación, al igual que con el modelo LSTM, se establece el número de iteraciones de entrenamiento mediante la variable *epochs*, fijada en 100 iteraciones completas sobre el conjunto de datos. Durante cada iteración, el modelo procesa las secuencias de entrenamiento contenidas en *dataloader* en forma de lotes. Para cada lote, los datos X se transfieren al dispositivo de cálculo y se realiza una propagación hacia adelante a través del modelo, obteniendo como salida la reconstrucción de la secuencia (x_{recon}) y los parámetros de la distribución latente (μ y $\logvar$).

Posteriormente, se calcula la pérdida utilizando la función *vae_loss*, y se ejecuta el proceso de retropropagación mediante *loss.backward()*, que permite calcular los gradientes necesarios para actualizar los parámetros del modelo. Finalmente, el optimizador aplica la actualización de los pesos mediante *optimizer.step()*. Durante el proceso se acumula la pérdida total de cada iteración en la variable *total_loss*, lo que permite monitorizar la evolución del entrenamiento. Asimismo, se registra el instante inicial del entrenamiento mediante la variable *start_time*, con el objetivo de calcular posteriormente el tiempo total necesario para entrenar el modelo, recogidos según el tipo de buque en la Tabla 3-4. Una vez finalizadas las iteraciones, se guarda el modelo en el archivo correspondiente para poder utilizarlo sin volver a entrenar. En el caso de este modelo, al contrario del modelo LSTM, no es necesario reducir el número de trayectorias para ningún tipo de buque.

Tipo de buque	Trayectorias de entrenamiento	Media de puntos	Tiempo de entrenamiento 1 epoch	Tiempo de entrenamiento 100 epoch	Tiempo de generación de 100 trayectorias
0-Carguero	7210	1581.69	53s	1h 28m 20s	3m 3s
1-Petrolero	22622	1378.16	48.32s	1h 20m 32s	4m 12s
2-Pesquero	31163	1187.08	35.18s	58m 38s	2m 57s
3-Pasajero	12336	698.63	8.75s	14m 35s	2m 33s
4-Remolcador	14844	755.97	12.63s	21m 03s	3m 42s

Tabla 3-4 Tiempos de entrenamiento del modelo SeqVAE según el tipo de buque

Una vez entrenado el modelo SeqVAE, se procede a su utilización para generar nuevas trayectorias sintéticas. En primer lugar, se inicializa el modelo y se cargan los parámetros previamente entrenados almacenados en el archivo correspondiente. Posteriormente, el modelo se establece en modo evaluación mediante *model.eval()* (Figura 3-16).

A continuación, se generan vectores latentes aleatorios z utilizando una distribución normal estándar. Estos vectores representan diferentes puntos dentro del espacio latente aprendido por el modelo y permiten producir nuevas muestras sintéticas. Mediante el decodificador del modelo, dichos vectores se transforman en secuencias sintéticas que mantienen las características aprendidas durante el entrenamiento.

```

model = SeqVAE().to(device)
model.load_state_dict(torch.load("modelo_trayectorias_pesqueros_VAE.pth"))
model.eval()

num_samples = 100
latent_dim = 16

z = torch.randn(num_samples, latent_dim).to(device)

with torch.no_grad():
    synthetic_scaled = model.decode(z, seq_len).cpu().numpy()

# Inverse scaling
n, t, f = synthetic_scaled.shape
synthetic_2d = synthetic_scaled.reshape(-1, f)
synthetic_real = scaler.inverse_transform(synthetic_2d)
synthetic_real = synthetic_real.reshape(n, t, f)

```

Figura 3-16 Generación de trayectorias sintéticas con SeqVAE

Las secuencias generadas se encuentran inicialmente en la escala normalizada utilizada durante el proceso de entrenamiento. Por este motivo, se aplica posteriormente el proceso inverso de normalización utilizando el objeto *scaler*, lo que permite recuperar los valores de las variables en su escala original. De esta manera, se obtienen finalmente las trayectorias sintéticas generadas por el modelo.

Una vez generadas las secuencias sintéticas mediante el modelo SeqVAE y recuperados los valores originales tras el proceso de desnormalización descrito anteriormente, es necesario transformar dichas secuencias en trayectorias geográficas completas (con latitud y longitud para poder representarlas). Como se ha explicado previamente, el modelo genera variables dinámicas del movimiento (seno y coseno del rumbo, velocidad y diferencia temporal) en lugar de coordenadas geográficas absolutas.

```

def recuperar_rumbo(sin_val, cos_val):
    ang = np.degrees(np.arctan2(sin_val, cos_val))
    return (ang + 360) % 360

```

Figura 3-17 Función *recuperar_rumbo* del modelo SeqVAE

En primer lugar, se reconstruye el rumbo a partir de su representación trigonométrica mediante la función *recuperar_rumbo* (Figura 3-17). Esta función permite obtener el ángulo correspondiente a partir de los valores de seno y coseno generados por el modelo, recuperando así la dirección real del movimiento.

A continuación, se calcula la posición geográfica siguiente utilizando la función *mover_punto*. Tal y como se ha descrito anteriormente, esta función estima una nueva posición geográfica a partir de una posición inicial, el rumbo de desplazamiento y la distancia recorrida sobre la superficie terrestre. La distancia se obtiene multiplicando la velocidad por el intervalo temporal entre observaciones.

Estos cálculos se aplican de forma iterativa mediante la función *generar_trayectoria*, que recorre cada uno de los pasos temporales de la secuencia sintética generada por el modelo. Partiendo de una posición inicial de referencia, en cada iteración se reconstruye el rumbo, se calcula la distancia recorrida y se obtiene la nueva posición geográfica, almacenando las coordenadas resultantes.

Finalmente, este procedimiento se aplica a todas las secuencias sintéticas generadas por el modelo, obteniendo así un conjunto de trayectorias sintéticas en coordenadas geográficas, que pueden ser utilizadas posteriormente para su análisis o visualización.

```
flat = synthetic_real.reshape(-1, 4)

sin_r = flat[:, 0]
cos_r = flat[:, 1]
vel = flat[:, 2]
dt = flat[:, 3]

heading_rad = np.arctan2(sin_r, cos_r)

heading_deg = np.degrees(heading_rad)

heading_deg = (heading_deg + 360) % 360

# crear dataframe
df_synthetic_real = pd.DataFrame({
    "rumbo": heading_deg,
    "velocidad": vel,
    "dt": dt
})
```

Figura 3-18 Recuperación del rumbo y estructuración de variables

En el bloque de la Figura 3-18 se reorganizan las variables generadas por el modelo con el objetivo de recuperar el rumbo y estructurar los datos en un formato más manejable. En primer lugar, las secuencias sintéticas se transforman en una estructura bidimensional, lo que permite tratar cada observación de forma individual.

A partir de esta estructura se separan las cuatro variables generadas por el modelo: el seno del rumbo, el coseno del rumbo, la velocidad y el intervalo temporal entre observaciones (dt). Posteriormente, se reconstruye el valor del rumbo a partir de su representación trigonométrica utilizando la función arcotangente. Este cálculo permite obtener inicialmente el rumbo en radianes, que posteriormente se convierte a grados.

Finalmente, el rumbo se normaliza para asegurar que sus valores se encuentren dentro del intervalo de 0° a 360° . Una vez recuperadas las variables en su forma interpretable, se organizan en una estructura tabular que contiene el rumbo, la velocidad y la diferencia temporal, facilitando así su utilización en las etapas posteriores del proceso.

Las trayectorias sintéticas generadas se organizan en una estructura tabular que recoge, para cada punto, el identificador de la trayectoria, el paso temporal y sus coordenadas geográficas. Posteriormente, estos datos se almacenan en un archivo en formato CSV, lo que facilita su almacenamiento y su utilización en etapas posteriores de análisis o visualización. Se guardan bajo el nombre, por ejemplo, *trayectorias_sinteticas_4.csv*, siendo el número 4 referente a los remolcadores.

Como paso final del proceso, se realiza una validación geográfica de las trayectorias sintéticas generadas con el objetivo de garantizar su plausibilidad física. Para ello se emplea información cartográfica de masas terrestres procedente de la base de datos *Natural Earth* [36]. Mediante esta información se comprueba si alguno de los puntos de las trayectorias generadas se encuentra sobre tierra firme, ya que las trayectorias AIS reales de los buques deben desarrollarse exclusivamente sobre el medio marítimo.

Las trayectorias que presentan intersecciones con zonas terrestres se descartan automáticamente, conservándose únicamente aquellas que se desarrollan completamente sobre el mar. Este filtrado permite

evitar trayectorias físicamente imposibles y garantiza que los datos sintéticos generados mantengan coherencia geográfica con el comportamiento real de los buques.

De este modo, el proceso descrito en este capítulo permite generar y reconstruir trayectorias sintéticas a partir de los datos originales, obteniendo finalmente un conjunto de trayectorias coherentes que podrán emplearse en los análisis presentados en el capítulo siguiente.

3.4 Generación de información temporal

Una vez generadas las trayectorias sintéticas mediante los dos diferentes modelos (LSTM y SeqVAE), las posiciones obtenidas únicamente contienen información espacial (latitud y longitud) y el índice de paso dentro de la trayectoria. Sin embargo, los datos AIS reales incluyen además una dimensión temporal asociada a cada posición, representada mediante un *timestamp*. Esta información temporal es esencial para poder calcular magnitudes dinámicas como la velocidad, la duración de la trayectoria o los intervalos de tiempo entre mensajes consecutivos.

Con el objetivo de evaluar el realismo de las trayectorias sintéticas generadas, resulta necesario comparar sus propiedades dinámicas con las observadas en los datos reales. En particular, uno de los indicadores más relevantes es la distribución de velocidades medias de las trayectorias. Para poder realizar dicha comparación, es imprescindible asignar una marca temporal a cada punto de las trayectorias sintéticas de manera coherente con los patrones temporales observados en los datos AIS reales.

El procedimiento implementado consiste en estimar primero la distribución empírica de velocidades a partir de las trayectorias reales y utilizar posteriormente dicha distribución para generar intervalos temporales en las trayectorias sintéticas. De este modo, se asegura que la dinámica temporal de las trayectorias simuladas sea estadísticamente comparable a la de los datos reales.

En primer lugar, se cargan los conjuntos de datos reales y sintéticos. El conjunto *df_trayectorias_limpias* contiene trayectorias obtenidas durante el proceso de preprocesamiento de los datos detallados en el apartado 3.2, mientras que *df_sintetico* contiene las trayectorias generadas por los diferentes modelos. Cada trayectoria sintética está compuesta por una secuencia de puntos identificados por un identificador de trayectoria (*traj_id*) y un índice temporal (*step*).

```
import pandas as pd
import numpy as np
df_sintetico = pd.read_csv("trayectorias_synth_0_15epoch.csv", low_memory=False)
df_trayectorias_limpias = pd.read_parquet("trayectorias_cargueros.parquet")

df_trayectorias_limpias["timestamp"] = pd.to_datetime(trayectorias_limpias["timestamp"], unit="s")

df_trayectorias_limpias = df_reales.sort_values(["id_trayectoria", "timestamp"])

df_trayectorias_limpias["dt"] = (
    df_reales.groupby("id_trayectoria")["timestamp"]
    .diff()
    .dt.total_seconds()
)

print(df_trayectorias_limpias["dt"].describe())
```

Figura 3-19 Carga de datos y conversión temporal para la generación de datos temporales

Los *timestamps* de los datos reales se almacenan en formato *Unix epoch* (segundos desde el 1 de enero de 1970). Para poder realizar operaciones temporales es necesario convertirlos a formato *datetime* como se muestra en la Figura 3-19. Posteriormente, los datos se ordenan por identificador de trayectoria y por tiempo para asegurar que las posiciones estén correctamente secuenciadas. Una vez ordenados los datos, se calcula la diferencia temporal entre posiciones consecutivas dentro de cada trayectoria (*dt*), este valor representa el intervalo de tiempo entre dos mensajes AIS consecutivos.

Los datos reales ya contienen una columna de la velocidad en nudos (*velocidad*). A partir de esta variable se calcula la distribución empírica de velocidades medias agrupadas en distintos rangos de 10 nudos (Figura 3-20).

```
bins = [0,10,20,30,40]

distribucion = (
    pd.cut(df_trayectorias_limpias["velocidad"], bins=bins)
    .value_counts(normalize=True)
    .sort_index()
)

print("\nDistribución real de velocidades:")
print(distribucion)

probs = distribucion.values
```

Figura 3-20 Cálculo de la distribución de velocidades medias de los datos reales

Esta distribución representa la probabilidad de que una trayectoria tenga una velocidad media dentro de cada intervalo definido.

A continuación, para generar *timestamps* coherentes con la velocidad, es necesario calcular la distancia entre posiciones consecutivas de las trayectorias sintéticas. Para ello se utiliza la fórmula de *Haversine* (Figura 3-21), que permite calcular la distancia sobre la superficie terrestre a partir de coordenadas geográficas. Esta función calcula la distancia entre dos puntos de la superficie terrestre considerando el radio medio de la Tierra. Posteriormente se calcula la diferencia de coordenadas entre cada punto y su posición anterior dentro de la misma trayectoria y se obtiene la distancia entre ambos puntos.

Una vez conocidas las distancias entre puntos, se generan los *timestamps* utilizando la distribución de velocidades reales (Figura 3-22). Para cada trayectoria sintética se selecciona aleatoriamente un rango de velocidad según la distribución observada en los datos reales. Dentro de ese rango se genera una velocidad media concreta. El intervalo temporal necesario para recorrer cada tramo se calcula mediante la relación fundamental entre distancia, velocidad y tiempo.

$$t = \frac{d}{v}$$

Donde t es el tiempo entre posiciones, d es la distancia recorrida y v la velocidad media asignada.

```
def haversine(lat1, lon1, lat2, lon2):

    R = 6371000

    lat1 = np.radians(lat1)
    lon1 = np.radians(lon1)
    lat2 = np.radians(lat2)
    lon2 = np.radians(lon2)

    dlat = lat2 - lat1
    dlon = lon2 - lon1

    a = np.sin(dlat/2)**2 + np.cos(lat1)*np.cos(lat2)*np.sin(dlon/2)**2
    c = 2*np.arcsin(np.sqrt(a))

    return R*c

# CALCULAR DISTANCIA ENTRE PUNTOS SINTÉTICOS

df_sintetico = df_sintetico.sort_values(["traj_id", "step"])

df_sintetico["lat_prev"] = df_sintetico.groupby("traj_id")["lat_pred"].shift()
df_sintetico["lon_prev"] = df_sintetico.groupby("traj_id")["lon_pred"].shift()

df_sintetico["distancia"] = haversine(
    df_sintetico["lat_prev"],
    df_sintetico["lon_prev"],
    df_sintetico["lat_pred"],
    df_sintetico["lon_pred"]
)

df_sintetico["distancia"] = df_sintetico["distancia"].fillna(0)
```

Figura 3-21 Función de cálculo de distancias de las trayectorias sintéticas

Posteriormente, se añade una pequeña perturbación aleatoria para simular la irregularidad temporal típica de los mensajes AIS. Finalmente, se genera la secuencia temporal acumulando los intervalos de tiempo calculados.

```
bins_vel = [(0,10),(10,20),(20,30),(30,40)]

df_sintetico["timestamp"] = None

for traj_id in df_sintetico["traj_id"].unique():

    mask = df_sintetico["traj_id"] == traj_id
    traj = df_sintetico.loc[mask]

    # elegir rango de velocidad según distribución real
    bin_idx = np.random.choice(len(bins_vel), p=probs)

    vmin, vmax = bins_vel[bin_idx]
    v_media = np.random.uniform(vmin, vmax)

    dist_nm = traj["distancia"].values / 1852

    dt_h = dist_nm / v_media
    dt_h = dt_h * np.random.normal(1,0.1,len(dt_h))

    dt_s = dt_h * 3600

    start_time = (
        pd.Timestamp("2024-01-01")
        + pd.Timedelta(minutes=np.random.randint(0,200000))
    )

    times = [start_time]

    for d in dt_s[1:]:
        times.append(times[-1] + pd.Timedelta(seconds=d))

    df_sintetico.loc[mask, "timestamp"] = times
```

Figura 3-22 Cálculo de los datos temporales para las trayectorias sintéticas

Una vez generados los datos temporales para todas las trayectorias, el nuevo conjunto de datos sintético se guarda en un archivo CSV bajo el nombre *trayectorias_con_tipodebuque.csv*.

4 VALIDACIÓN DE LOS RESULTADOS

En este capítulo se evalúan las trayectorias sintéticas generadas por los modelos LSTM y SeqVAE, comparándolas con las trayectorias reales. La validación se aborda desde tres perspectivas complementarias: un análisis de características generales, una inspección visual de las trayectorias, y una comparación detallada de métricas estadísticas.

El objetivo es determinar hasta qué punto los modelos son capaces de reproducir tanto la estructura global del tráfico marítimo como las características específicas de cada tipo de buque. Asimismo, se analiza si los modelos logran capturar las diferencias entre categorías, como trayectorias de largo recorrido en cargueros y petroleros, patrones más locales en pesqueros o recorridos cortos en remolcadores.

4.1 Características generales

En este apartado se lleva a cabo una evaluación global de las trayectorias sintéticas generadas mediante los modelos LSTM y SeqVAE, comparándolas con las trayectorias reales obtenidas a partir de datos AIS. Se adopta un enfoque basado en dos aspectos principales: por un lado, el análisis de características estadísticas agregadas de las trayectorias y, por otro, una inspección visual de los recorridos generados.

El objetivo de este análisis es evaluar de forma conjunta la coherencia global de las trayectorias sintéticas, determinando hasta qué punto los modelos son capaces de reproducir no solo las distribuciones individuales, sino también el comportamiento general del tráfico marítimo en términos de magnitudes medias y patrones espaciales.

Para la inspección visual de las trayectorias, se ha optado por representar un subconjunto reducido de trayectorias con el fin de evitar la sobrecarga visual en los mapas. En concreto, se muestran 50 trayectorias reales y 50 trayectorias generadas mediante el modelo LSTM, lo que permite una comparación clara sin pérdida de legibilidad. En el caso del modelo SeqVAE, dado que las trayectorias generadas parten de puntos iniciales aleatorios procedentes de trayectorias reales, se representa igualmente un subconjunto de trayectorias, evitando así la dispersión excesiva y facilitando la interpretación de los resultados.

A partir de los valores recogidos en la Tabla 4-1, la Tabla 4-2 y la Tabla 4-3, se comparan las velocidades medias, distancias recorridas y tiempos medios de las trayectorias reales con las generadas por ambos modelos.

Tipo de buque	Velocidad media (nudos)	Distancia media (millas náuticas)	Tiempo medio (horas)
0-Carguero	14.94	687.52	58.87
1-Petrolero	14.61	637.36	53.13
2-Pesquero	6.96	252.02	47.27
3-Pasajero	11.38	194.09	30.36
4-Remolcador	3.77	79.67	40.11

Tabla 4-1 Estadísticas generales de las trayectorias reales

En términos generales, el modelo **LSTM** tiende a generar trayectorias con **distancias medias** significativamente mayores que las observadas en los datos reales, lo que se traduce también en tiempos medios considerablemente más elevados. Sin embargo, las velocidades medias generadas se mantienen relativamente próximas a las reales, lo que indica que el modelo conserva adecuadamente la relación entre desplazamiento y tiempo.

Tipo de buque	Velocidad media (nudos)	Distancia media (millas náuticas)	Tiempo medio (horas)
0-Carguero	12.52	975.15	443.30
1-Petrolero	13.71	1129.08	236.55
2-Pesquero	9.88	1073.40	134.92
3-Pasajero	12.31	550.43	55.72
4-Remolcador	8.67	550.43	72.58

Tabla 4-2 Estadísticas generales de las trayectorias sintéticas LSTM

Por otro lado, el modelo **SeqVAE** presenta un comportamiento opuesto, generando trayectorias con distancias y tiempos medios considerablemente menores que los observados en los datos reales. Este resultado refleja la tendencia del modelo a producir trayectorias más cortas. A pesar de ello, las velocidades medias se mantienen en rangos similares a los reales, lo que sugiere que el modelo es capaz de capturar parcialmente la dinámica del movimiento, aunque no su escala espacial.

En conjunto, este análisis pone de manifiesto que ambos modelos presentan comportamientos diferenciados: mientras que el modelo LSTM tiende a sobrestimar la extensión de las trayectorias, el modelo SeqVAE tiende a infraestimarla, situándose los datos reales en un punto intermedio entre ambos enfoques.

Tipo de buque	Velocidad media (nudos)	Distancia media (millas náuticas)	Tiempo medio (horas)
0-Carguero	14.14	82.02	9.87
1-Petrolero	12.31	112.93	22.12
2-Pesquero	6.79	68.44	50.52
3-Pasajero	11.18	58.15	22.04
4-Remolcador	6.44	45.88	17.09

Tabla 4-3 Estadísticas generales de las trayectorias sintéticas SeqVAE

4.1.1 Cargueros

A continuación, se realiza una **inspección visual de las trayectorias** (Figura 4-1, Figura 4-2 y Figura 4-3) para el caso de los buques cargueros, comparando un subconjunto de trayectorias reales con las generadas por el modelo LSTM y SeqVAE.

En la representación de las trayectorias reales se observa una distribución geográfica amplia, con rutas que conectan distintas regiones del globo y que reflejan patrones realistas del tráfico marítimo internacional. Las trayectorias presentan orientaciones diversas y cubren tanto rutas de larga distancia como recorridos más regionales.

En el caso de las trayectorias generadas mediante el modelo **LSTM**, se aprecia una estructura global coherente, con recorridos que siguen trayectorias suaves y continuas. No obstante, se observa una mayor concentración en determinadas zonas geográficas, así como una menor dispersión en comparación con los datos reales. Además, algunas trayectorias presentan recorridos más extensos, lo que es consistente con la tendencia del modelo a generar distancias mayores.

En el caso de las trayectorias generadas mediante el modelo **SeqVAE**, las trayectorias generadas presentan recorridos relativamente cortos y con una extensión limitada. No obstante, las trayectorias mantienen una estructura coherente, con recorridos continuos y sin discontinuidades, lo que indica que el modelo es capaz de generar secuencias espaciales consistentes.

En conjunto, la inspección visual muestra que el modelo SeqVAE genera trayectorias plausibles a nivel local, pero con una extensión limitada.



Figura 4-1 Representación de 50 trayectorias reales de Cargueros

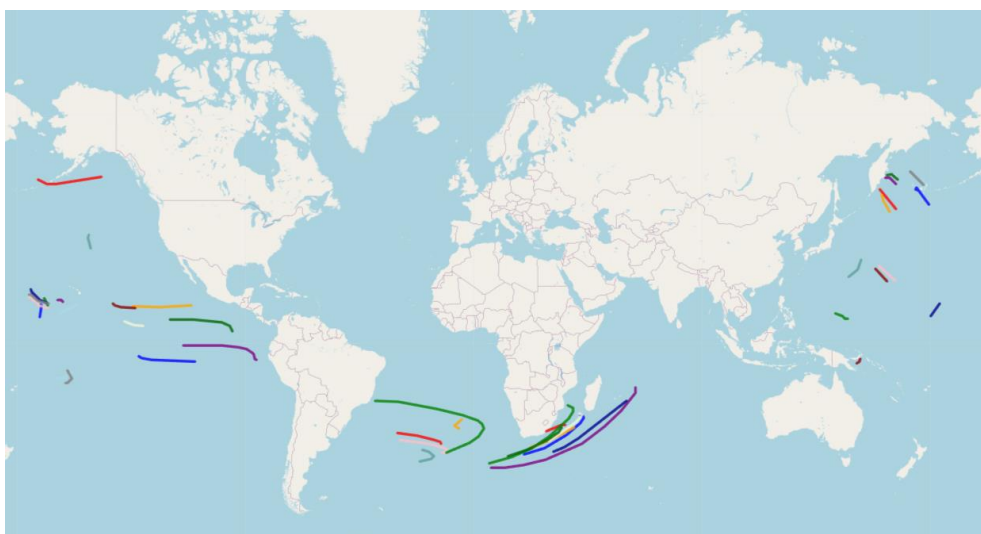


Figura 4-2 Representación de 50 trayectorias sintéticas generadas con LSTM para Cargueros

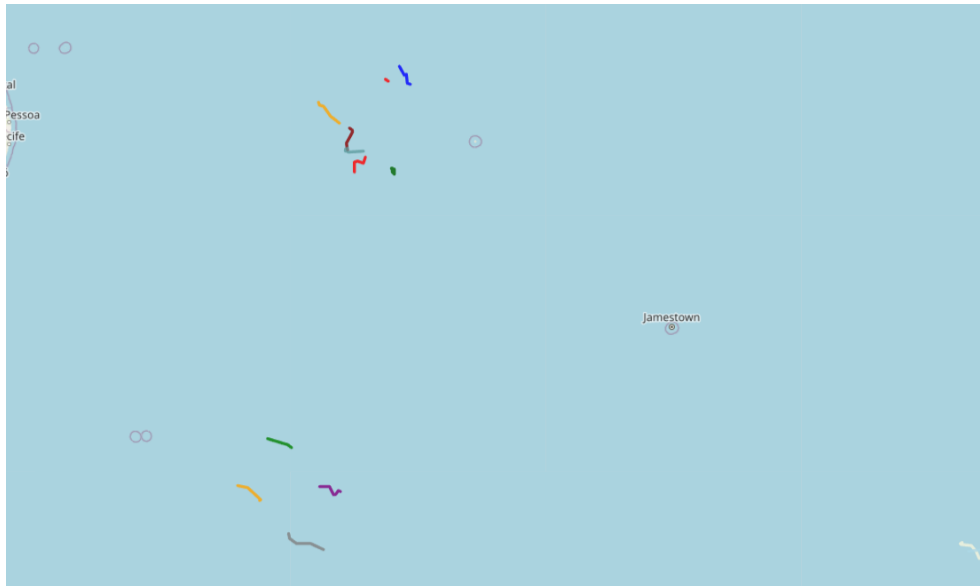


Figura 4-3 Representación de una selección de trayectorias sintéticas generadas con SeqVAE para Cargueros

4.1.2 Petroleros

En este apartado se realiza una evaluación visual de las trayectorias correspondientes a los buques petroleros (Figura4-4, Figura4-5 y Figura 4-6), comparando un subconjunto de trayectorias reales con las generadas mediante los modelos LSTM y SeqVAE.

De forma similar a lo observado en el caso de los cargueros, las trayectorias reales presentan una distribución geográfica amplia y diversa, con recorridos que conectan distintas regiones del mundo y que reflejan rutas marítimas de media y larga distancia. Las trayectorias muestran una elevada variabilidad en orientación y longitud, manteniéndose en zonas oceánicas y evitando áreas terrestres, lo que evidencia un comportamiento coherente con la navegación real.

En las trayectorias generadas mediante el modelo **LSTM** se aprecia un comportamiento globalmente consistente con el descrito previamente para los cargueros. Los recorridos son suaves y continuos, sin discontinuidades apreciables, y presentan una distribución espacial relativamente amplia, con diversidad en las direcciones de desplazamiento. No obstante, al igual que en el caso anterior, se observa cierta tendencia a generar trayectorias más extensas, así como una ligera concentración en determinadas zonas geográficas en comparación con los datos reales.

Por otro lado, las trayectorias generadas mediante el modelo **SeqVAE** mantienen un patrón caracterizado por recorridos de menor extensión en comparación con las trayectorias reales. Al partir de puntos iniciales aleatorios procedentes de trayectorias reales, las trayectorias sintéticas se distribuyen de forma más coherente en el espacio geográfico, evitando concentraciones en una única zona de origen.

En conjunto, y en línea con lo observado para los cargueros, la inspección visual muestra que el modelo LSTM genera trayectorias más realistas en términos de distribución espacial y diversidad, mientras que el modelo SeqVAE presenta limitaciones en la dispersión geográfica y en la coherencia con el entorno marítimo.



Figura 4-4 Representación de 50 trayectorias reales de Petroleros

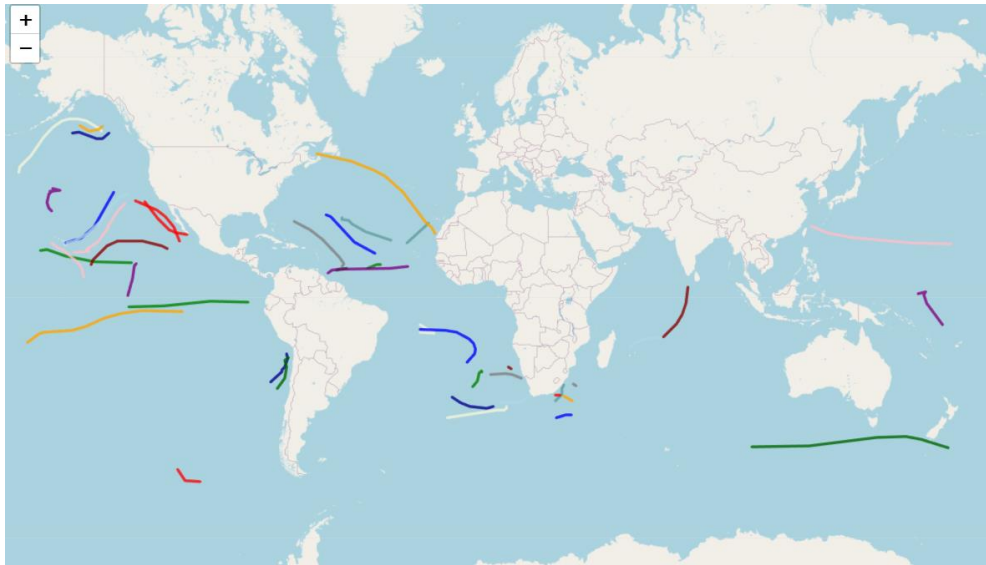


Figura 4-5 Representación de 50 trayectorias sintéticas generadas con LSTM para Petroleros



Figura 4-6 Representación de una selección de trayectorias sintéticas generadas con SeqVAE para Petroleros

4.1.3 Pesqueros

En este apartado se realiza una evaluación visual de las trayectorias correspondientes a los buques pesqueros (Figura 4-7, Figura 4-8 y Figura 4-9), comparando las trayectorias reales con las generadas mediante los modelos LSTM y SeqVAE.

En las trayectorias reales se observa un comportamiento claramente diferenciado respecto a los tipos de buque analizados anteriormente. A diferencia de cargueros y petroleros, las trayectorias de los pesqueros presentan una menor extensión espacial y una mayor concentración en zonas concretas, lo que refleja su actividad en áreas de pesca determinadas. Los recorridos tienden a ser más cortos y, en muchos casos, muestran cambios de dirección más frecuentes, asociados a patrones de operación locales.

Las trayectorias generadas mediante el modelo **LSTM** mantienen en parte estas características, mostrando recorridos de menor longitud en comparación con otros tipos de buques y una mayor concentración en determinadas zonas. No obstante, se observa que algunas trayectorias presentan una extensión superior a la esperada o patrones más lineales, alejándose del comportamiento más irregular característico de los pesqueros. Aun así, el modelo es capaz de generar trayectorias continuas y espacialmente coherentes, aunque estas sean visualmente diferentes de las reales.

Por otro lado, las trayectorias generadas mediante el modelo **SeqVAE** presentan una menor variabilidad. Al partir de puntos iniciales aleatorios procedentes de trayectorias reales, las trayectorias sintéticas se sitúan en zonas coherentes con la actividad pesquera. Los recorridos son, en general, más cortos y con menor diversidad en cuanto a orientación y forma, tendiendo a mantenerse dentro de áreas geográficas más limitadas.

Este comportamiento es consistente con la naturaleza de las trayectorias reales de pesqueros, caracterizadas por movimientos más locales. No obstante, el modelo muestra una menor capacidad para reproducir la variabilidad observada en los datos reales, generando patrones más homogéneos.

En conjunto, y en línea con lo observado en los apartados anteriores, el modelo LSTM ofrece una representación más cercana al comportamiento real, mientras que el modelo SeqVAE genera trayectorias más homogéneas y con menor diversidad, lo que limita su capacidad para reproducir patrones de movimiento más complejos.



Figura 4-7 Representación de 50 trayectorias reales de Pesqueros

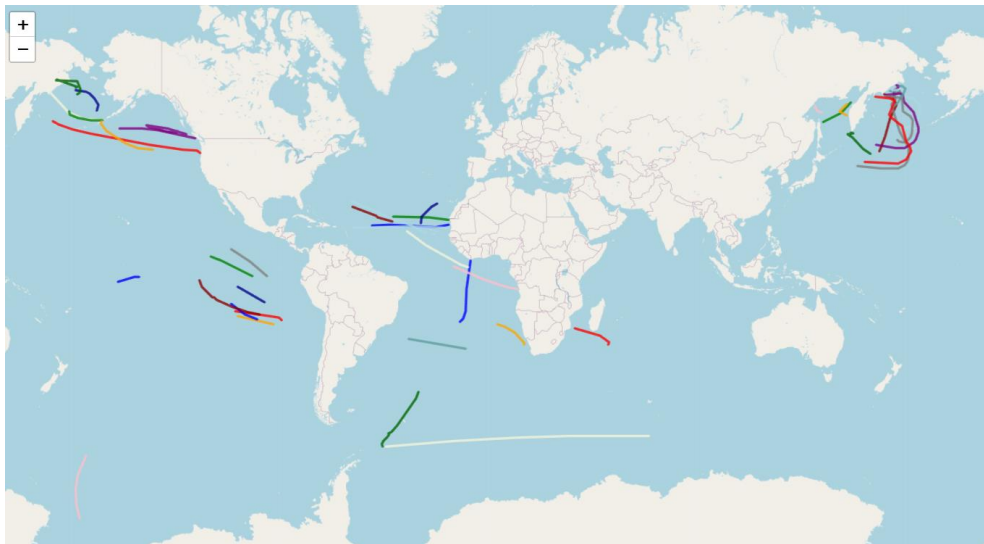


Figura 4-8 Representación de 50 trayectorias sintéticas generadas con LSTM para Pesqueros

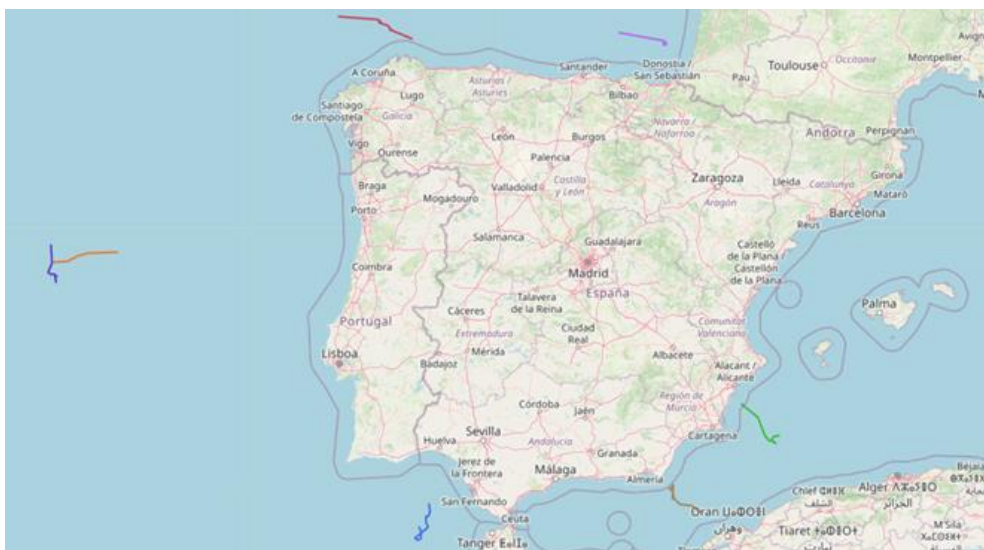


Figura 4-9 Representación de una selección de trayectorias sintéticas generadas con SeqVAE para Pesqueros

4.1.4 Pasajeros

En este apartado se realiza una evaluación visual de las trayectorias correspondientes a los buques de pasajeros (Figura 4-10, Figura 4-11 y Figura 4-12), comparando las trayectorias reales con las generadas mediante los modelos LSTM y SeqVAE.

Las trayectorias reales muestran, en general, recorridos de menor extensión que los observados en cargueros y petroleros, con rutas más localizadas y asociadas a conexiones regionales entre puertos. Se aprecia una distribución geográfica variada, aunque con trayectorias relativamente cortas y directas.

Las trayectorias generadas mediante el modelo **LSTM** presentan un comportamiento coherente, manteniendo recorridos continuos y una dispersión espacial razonable. No obstante, en algunos casos se observan trayectorias más extensas de lo esperado, alejándose parcialmente del carácter más local de este tipo de tráfico.

Por su parte, las trayectorias generadas mediante el modelo **SeqVAE** muestran nuevamente recorridos más limitados en extensión. Al partir de puntos iniciales aleatorios procedentes de trayectorias reales, las trayectorias sintéticas se sitúan en zonas coherentes con el tráfico marítimo de pasajeros. No obstante, los recorridos tienden a ser más homogéneos y con menor diversidad en cuanto a orientación y forma.

Este comportamiento indica que, aunque el modelo es capaz de generar trayectorias espacialmente coherentes, presenta limitaciones para reproducir la diversidad de rutas observada en los datos reales, especialmente en trayectorias que conectan diferentes regiones.

En conjunto, el modelo LSTM reproduce de forma más adecuada la diversidad espacial de las trayectorias, mientras que el modelo SeqVAE genera recorridos más homogéneos y con menor capacidad de generalización.



Figura 4-10 Representación de 50 trayectorias reales de Pasajeros

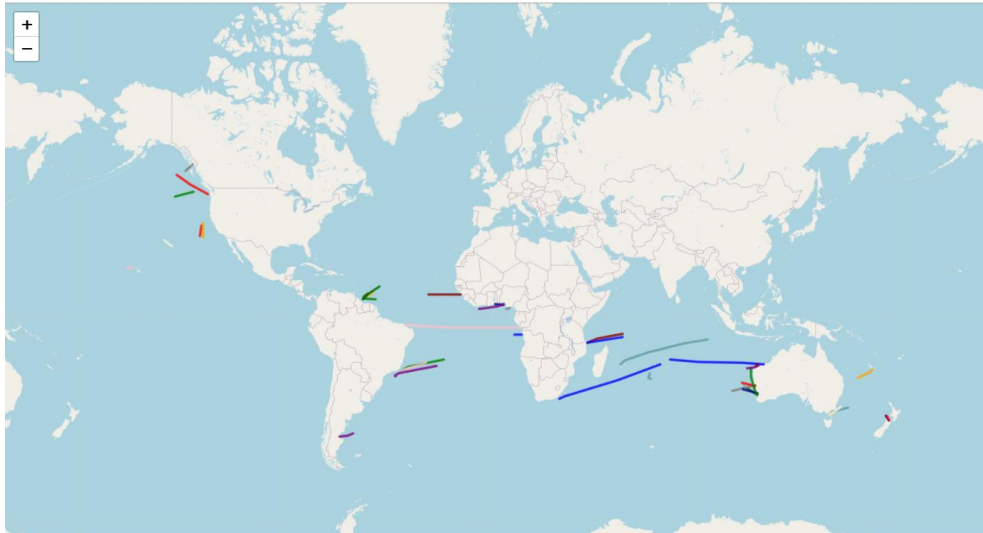


Figura 4-11 Representación de 50 trayectorias sintéticas generadas con LSTM para Pasajeros

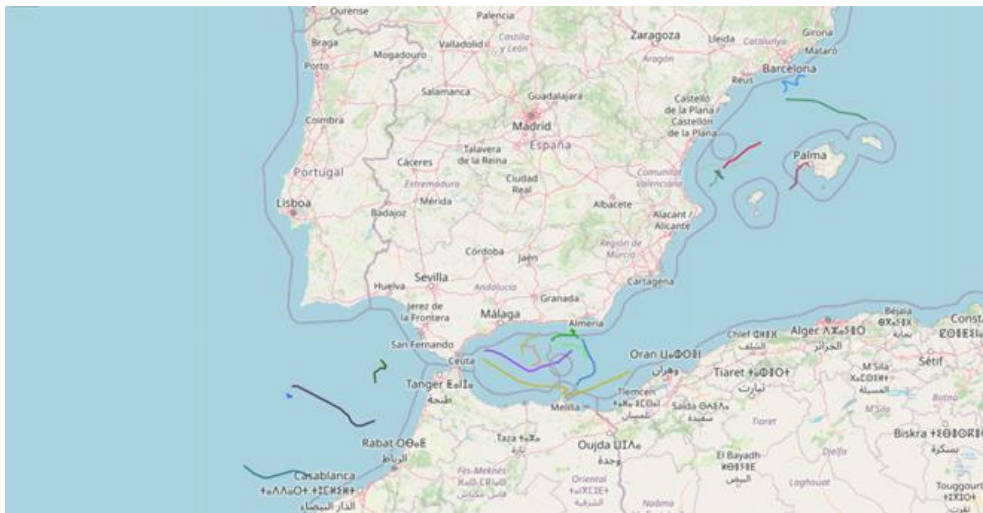


Figura 4-12 Representación de una selección de trayectorias sintéticas generadas con SeqVAE para Pasajeros

4.1.5 Remolcadores

En este apartado se realiza una evaluación visual de las trayectorias correspondientes a los remolcadores (Figura 4-13, Figura 4-14 y Figura 4-15), comparando las trayectorias reales con las generadas mediante los modelos LSTM y SeqVAE.

Las trayectorias reales presentan, en general, recorridos muy cortos y localizados, lo que refleja el ámbito de operación típico de este tipo de buques, principalmente en zonas portuarias o áreas cercanas a la costa. La dispersión geográfica es limitada y los trayectos muestran una clara naturaleza local.

Las trayectorias generadas mediante el modelo **LSTM** reproducen parcialmente este comportamiento, generando recorridos relativamente cortos y manteniendo una cierta coherencia espacial. No obstante, en algunos casos se observan trayectorias con mayor extensión de la esperada, alejándose del carácter estrictamente local de los remolcadores.

Por su parte, las trayectorias generadas mediante el modelo **SeqVAE** muestran una menor variabilidad en los recorridos. Los trayectos generados son, en general, más cortos y presentan una menor diversidad en comparación con los datos reales.

Este comportamiento es coherente con los resultados cuantitativos observados previamente, donde se evidenciaba una infraestimación de las distancias recorridas por parte del modelo. Asimismo, y en

línea con lo analizado en otros tipos de buque, el SeqVAE tiende a generar patrones más homogéneos, mostrando una capacidad limitada para reproducir la variabilidad espacial presente en los datos reales.

En términos globales, el modelo LSTM ofrece una representación más cercana al comportamiento real, mientras que el modelo SeqVAE presenta una menor capacidad para capturar la variabilidad y el carácter local de las trayectorias de los remolcadores.



Figura 4-13 Representación de 50 trayectorias reales de Remolcadores

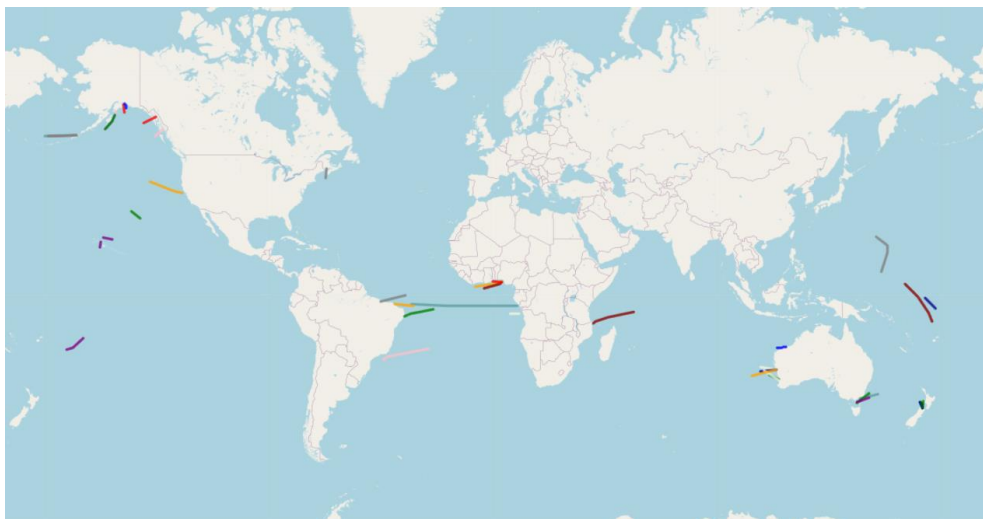


Figura 4-14 Representación de 50 trayectorias sintéticas generadas con LSTM para Remolcadores



Figura 4-15 Representación de una selección de trayectorias sintéticas generadas con SeqVAE para Remolcadores

4.2 Comparación de trayectorias generadas por LSTM y datos de trayectorias reales

En este apartado se comparan las trayectorias generadas mediante el modelo **LSTM** con los datos de trayectorias reales, analizando distintas características de los buques explicadas en el apartado 3.2.3.

En este apartado se ha optado por centrar el análisis en los tipos de buque cargueros y pesqueros, al considerarse los más representativos de los errores y comportamientos observados en el modelo LSTM. Estos tipos permiten ilustrar de forma clara tanto las desviaciones en trayectorias de largo recorrido como en trayectorias de carácter más local. El análisis correspondiente al resto de tipos de buque, incluyendo petroleros, pasajeros y remolcadores, se recoge en el Anexo III, con el objetivo de no sobrecargar el contenido principal y mantener la claridad en la exposición de resultados.

En términos generales, y en coherencia con lo observado en la inspección visual del apartado 4.1, el modelo LSTM es capaz de reproducir adecuadamente la estructura global de las trayectorias, manteniendo distribuciones realistas en velocidad, duración y distancia. No obstante, los resultados muestran, de forma recurrente, una tendencia a generar trayectorias más largas, lo que se traduce en mayores distancias y tiempos medios.

Este comportamiento se mantiene de forma consistente entre los distintos tipos de buque, aunque con matices relevantes. En el caso de cargueros y petroleros, el modelo reproduce razonablemente bien las trayectorias de largo recorrido, aunque con cierta sobreestimación. Para pesqueros y remolcadores, cuyas trayectorias reales son más cortas y locales, el modelo también captura esta naturaleza, aunque introduce en ocasiones recorridos más extensos de lo esperado.

En conjunto, el modelo LSTM demuestra una buena capacidad para diferenciar entre tipos de trayectorias, adaptando parcialmente su comportamiento a cada categoría de buque, aunque con una ligera tendencia sistemática a ampliar la escala de los recorridos.

4.2.1 Cargueros

En primer lugar, se analiza el comportamiento del modelo para el caso de los buques cargueros, comparando las características de las trayectorias reales con las generadas mediante el modelo LSTM.

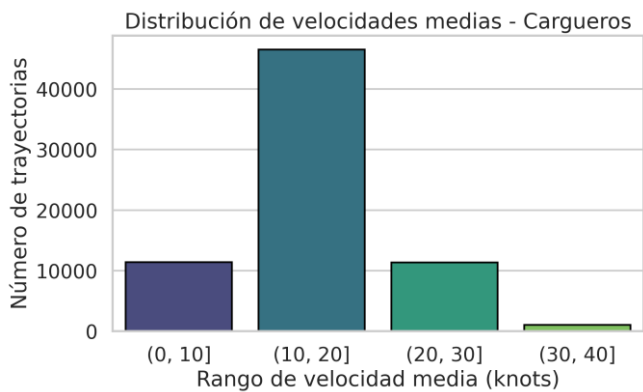


Figura 4-16 Distribución velocidades reales Cargueros

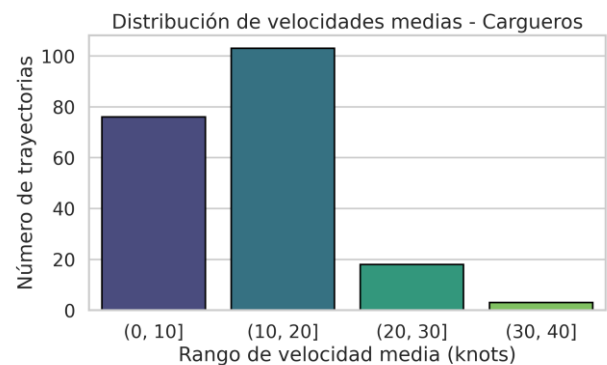


Figura 4-17 Distribución velocidades LSTM Cargueros

En el caso de los buques cargueros, caracterizados por trayectorias de media y larga distancia con carácter interregional o intercontinental, el modelo LSTM es capaz de reproducir de forma razonable la estructura general de los recorridos. Tal y como se ha observado en el apartado 4.1, así como en las representaciones visuales correspondientes, las trayectorias generadas presentan una distribución espacial coherente y mantienen la naturaleza global de este tipo de tráfico.

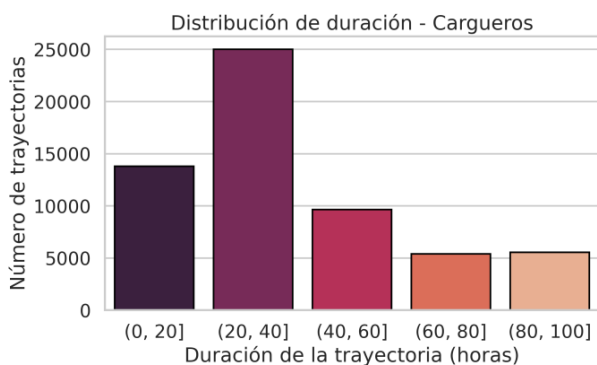


Figura 4-18 Distribución tiempo reales Cargueros

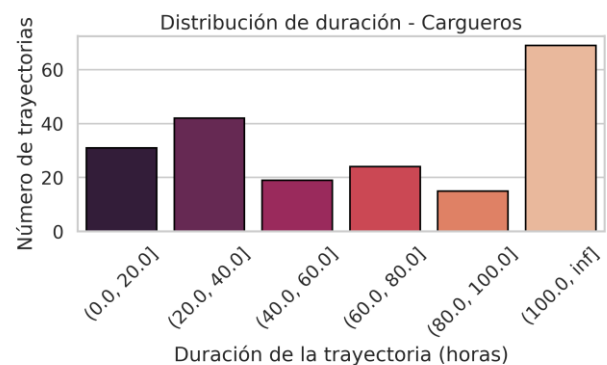


Figura 4-19 Distribución tiempo LSTM Cargueros

Desde el punto de vista estadístico, y como se aprecia en las gráficas de **distribución de velocidades, duración y distancia** (Figura 4-16, Figura 4-17, Figura 4-18, Figura 4-19, Figura 4-20 y Figura 4-21), el modelo reproduce adecuadamente los rangos predominantes de velocidad, así como la forma general de las distribuciones. No obstante, se observa de forma consistente una sobrestimación en

la distancia recorrida y en la duración de las trayectorias, lo que se refleja en un desplazamiento de las distribuciones hacia valores más elevados en comparación con los datos reales.

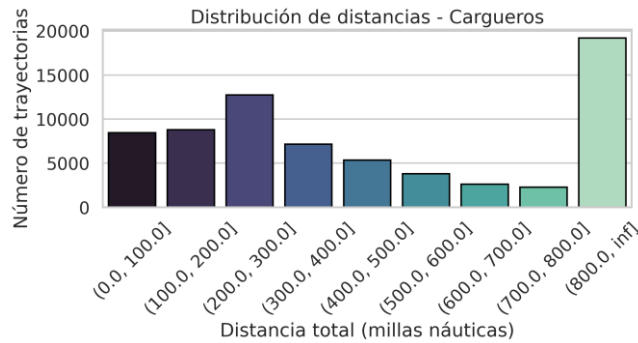


Figura 4-20 Distribución distancias reales Cargueros

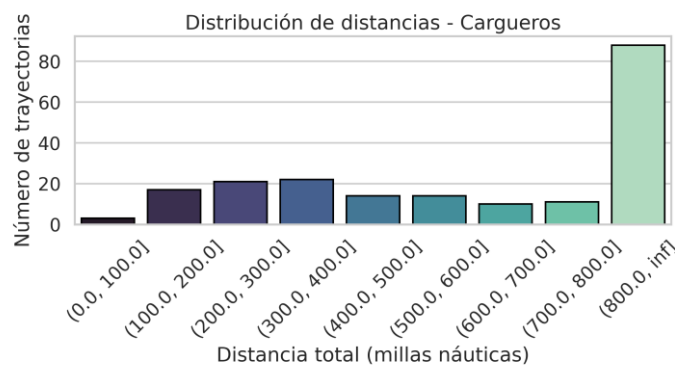


Figura 4-21 Distribución distancias LSTM Cargueros

Este comportamiento indica que, aunque el modelo es capaz de capturar correctamente el tipo de trayectoria asociado a los cargueros (trayectorias largas y continuas), tiende a amplificar la escala de los recorridos, generando trayectorias más extensas de lo esperado

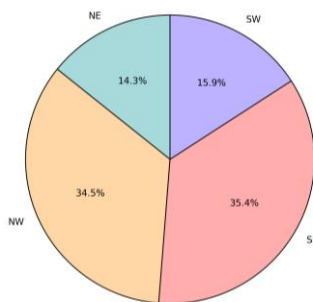


Figura 4-22 Distribución geográfica reales Cargueros

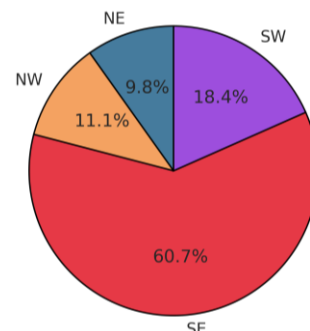


Figura 4-23 Distribución geográfica LSTM Cargueros

Por último, se analiza la **distribución geográfica de las trayectorias** (Figura 4-22 y Figura 4-23), clasificada según el cuadrante predominante de desplazamiento. En los datos reales los resultados muestran una distribución relativamente equilibrada entre los distintos cuadrantes, con cierta predominancia de trayectorias en el noroeste (NW) y el sureste (SE). Sin embargo, en las trayectorias generadas por el modelo LSTM se aprecia una mayor concentración en el cuadrante sureste (SE), mientras que el número de trayectorias en el cuadrante noreste (NE) es notablemente inferior.

Esta diferencia puede explicarse, en parte, por las características geográficas de las zonas incluidas en dicho cuadrante. El área correspondiente al cuadrante NE engloba regiones como el Mediterráneo oriental, el estrecho de Suez o rutas próximas a Europa y Oriente Medio, donde las rutas marítimas

discurren por corredores relativamente estrechos y cercanos a la costa. Durante el proceso de generación de trayectorias sintéticas se aplica un filtrado que descarta aquellas trayectorias que presentan puntos situados sobre tierra firme. Debido a la proximidad de las rutas marítimas a zonas costeras en este cuadrante, es más probable que algunas trayectorias generadas incluyan puntos no válidos, lo que provoca que dichas trayectorias sean eliminadas durante el proceso de validación. Como consecuencia, el número final de trayectorias sintéticas válidas en este cuadrante resulta menor que en otras regiones más abiertas.

En términos globales, los resultados muestran que el modelo LSTM diferencia adecuadamente este tipo de buque frente a otros con trayectorias más locales, reproduciendo su carácter de navegación de largo recorrido, aunque con una tendencia a generar trayectorias más extensas y con una ligera reducción en la dispersión geográfica.

4.2.2 Pesqueros

A continuación, se analiza el comportamiento del modelo para el caso de los buques pesqueros, caracterizados por trayectorias cortas, locales y con cambios de dirección frecuentes, asociadas a la actividad en zonas de pesca.

En línea con el análisis realizado en el apartado 4.1, las trayectorias generadas mediante el modelo LSTM reflejan en parte este carácter local, mostrando una menor extensión espacial que en otros tipos de buque, aunque sin ajustarse completamente a la irregularidad propia de este tipo de navegación.

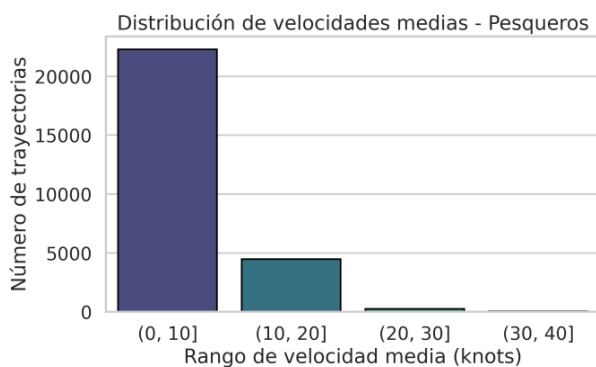


**Figura 4-24 Distribución velocidades reales
Pesqueros**

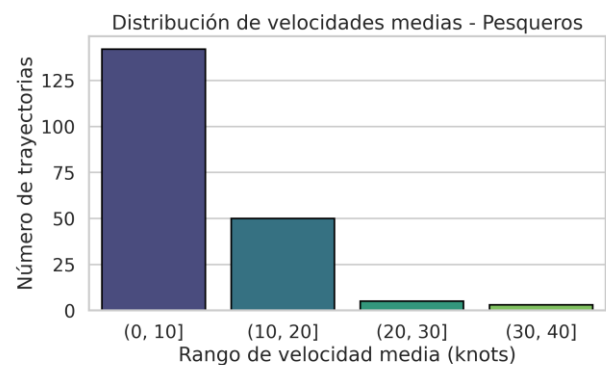


**Figura 4-25 Distribución velocidades LSTM
Pesqueros**

En las gráficas de distribución de velocidades (Figura 4-24 y 25) se observa que el modelo captura adecuadamente la predominancia de velocidades bajas, concentradas en el intervalo de 0 a 10 nudos, lo que constituye uno de los rasgos más representativos de los pesqueros. Sin embargo, al analizar las distribuciones de duración y distancia (Figura 4-26, Figura 4-27, Figura 4-28 y Figura 4-29), se aprecia una mayor presencia de trayectorias en intervalos superiores a los observados en los datos reales.

Esta desviación se traduce en trayectorias más prolongadas y extensas, alejándose parcialmente del carácter local de este tipo de buque. Como en casos anteriores, este comportamiento puede estar influido por la reconstrucción probabilística de la componente temporal.

En cuanto a la **distribución geográfica** (Figura 4-30 y Figura 4-31), las trayectorias generadas mantienen presencia en los distintos cuadrantes, aunque con ligeras variaciones respecto a los datos reales. Estas diferencias están condicionadas por el proceso de validación, especialmente relevante en este tipo de buque debido a su proximidad a la costa, lo que aumenta la probabilidad de descartar trayectorias no válidas.

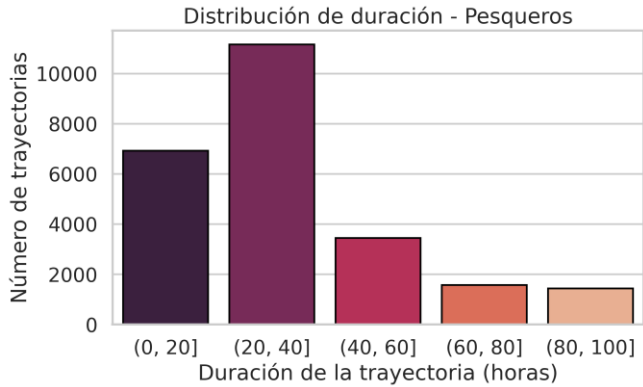


Figura 4-26 Distribución tiempo reales Pesqueros

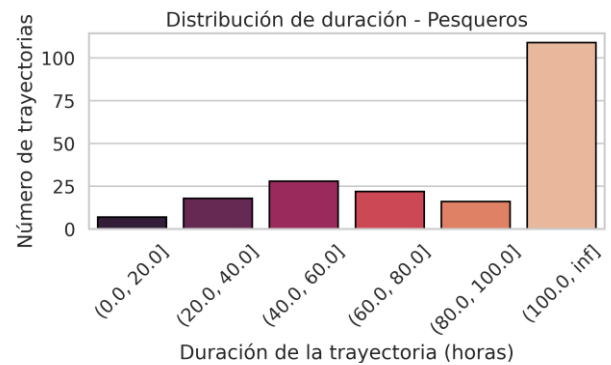


Figura 4-27 Distribución tiempo LSTM Pesqueros

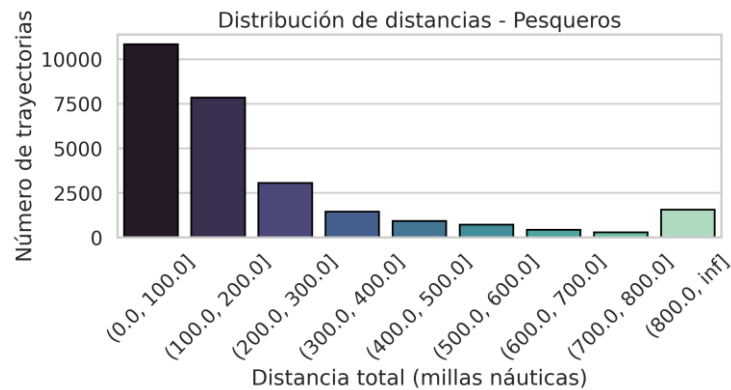


Figura 4-28 Distribución distancias reales Pesqueros

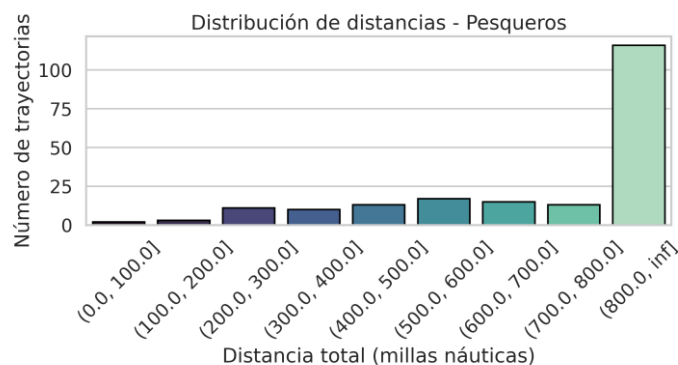


Figura 4-29 Distribución distancias LSTM Pesqueros

En conjunto, el modelo LSTM capta correctamente la naturaleza de baja velocidad y ámbito local de los pesqueros, aunque presenta limitaciones para reproducir su variabilidad espacial y tiende a generar trayectorias algo más extensas de lo esperado.

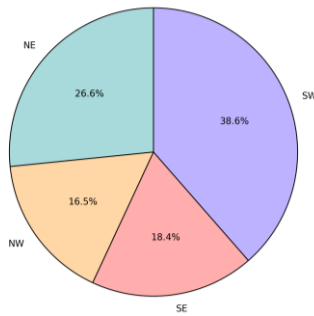


Figura 4-30 Distribución geográfica reales Pesqueros

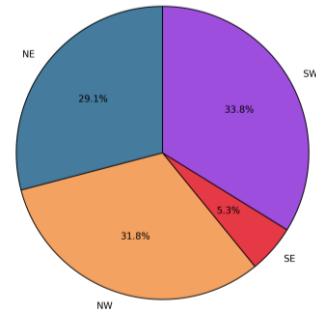


Figura 4-31 Distribución geográfica LSTM Pesqueros

4.3 Comparación de trayectorias generadas con SeqVAE y datos de trayectorias reales

En este apartado se analizan las trayectorias generadas mediante el modelo SeqVAE, comparándolas con los datos de trayectorias reales.

Tal y como se ha observado en el análisis visual del apartado 4.1, en este caso no se considera adecuado evaluar la distribución geográfica de las trayectorias generadas. Esto se debe a que, durante el proceso de generación, el modelo SeqVAE incorpora de forma intencionada puntos iniciales procedentes de trayectorias reales, lo que condiciona directamente la localización de las trayectorias sintéticas.

Como consecuencia, la distribución espacial de las trayectorias generadas no es independiente del conjunto de datos original, sino que reproduce en gran medida la distribución geográfica de las trayectorias reales. Por ello, un análisis de la dispersión o cobertura geográfica no aportaría información relevante sobre la capacidad generativa del modelo, ya que estaría reflejando principalmente la distribución de los datos de partida.

En consecuencia, el análisis de este apartado se centrará en otras características de las trayectorias, como su distribución de velocidades, duración y distancia, que permiten evaluar de forma más adecuada el rendimiento del modelo SeqVAE.

4.3.1 Cargueros

A continuación, se analiza el comportamiento del modelo SeqVAE para el caso de los buques cargueros, caracterizados por trayectorias de larga distancia y carácter global.

Tal y como se ha observado en el apartado 4.1, las trayectorias generadas presentan una cobertura espacial limitada y una menor dispersión en comparación con los datos reales, lo que condiciona la capacidad del modelo para representar recorridos interregionales o intercontinentales.

Desde el punto de vista estadístico, y como se aprecia en las **gráficas de distribución de velocidades, duración y distancia** (Figura 4-32, Figura 4-33, Figura 4-34, Figura 4-35, Figura 4-36 y Figura 4-37), las trayectorias generadas muestran una clara concentración en los intervalos de menor duración y distancia. Esto indica que el modelo tiende a generar trayectorias más cortas, alejándose del comportamiento característico de los cargueros.

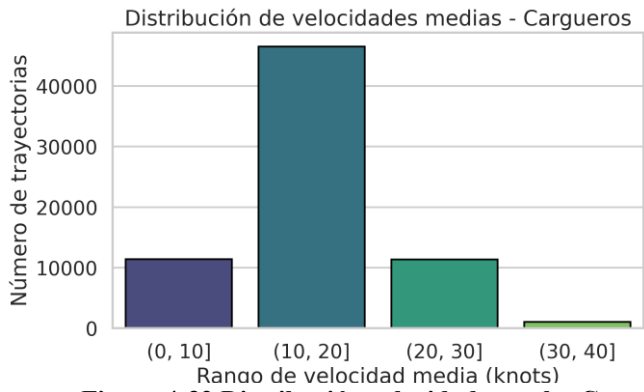


Figura 4-32 Distribución velocidades reales Cargueros

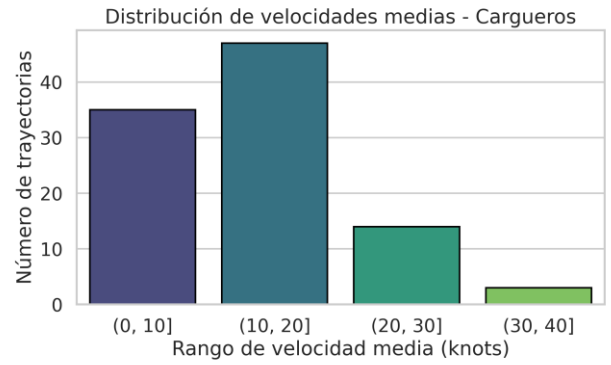


Figura 4-33 Distribución velocidades SeqVAE Cargueros

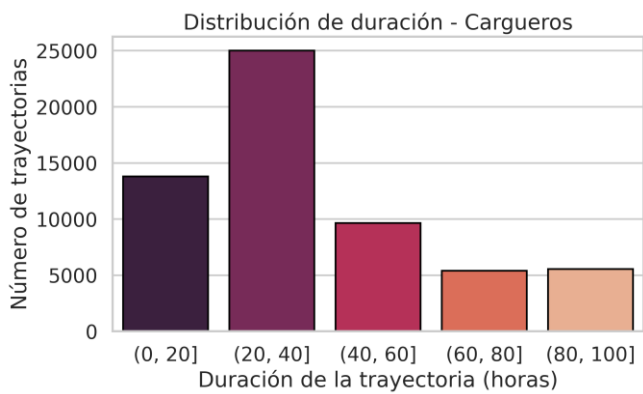


Figura 4-34 Distribución tiempo reales Cargueros

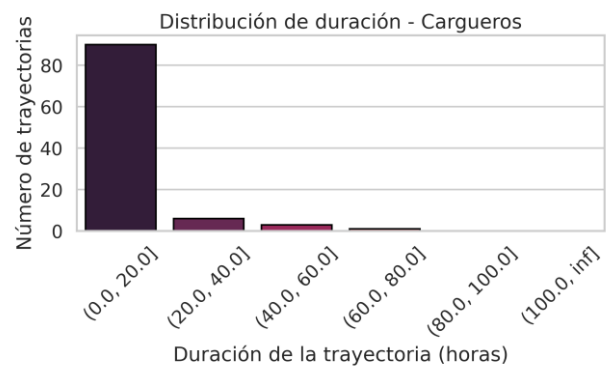


Figura 4-35 Distribución tiempo SeqVAE Cargueros

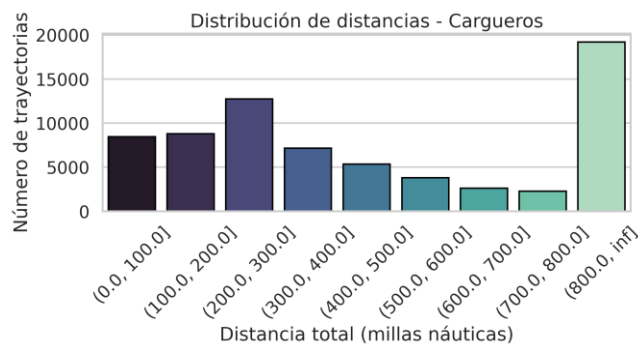


Figura 4-36 Distribución distancias reales Cargueros

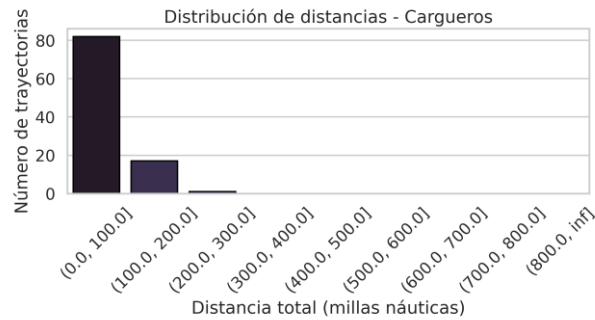


Figura 4-37 Distribución distancias SeqVAE Cargueros

En conjunto, los resultados muestran que el modelo SeqVAE presenta dificultades para capturar el patrón de trayectorias de larga distancia propio de los cargueros, generando recorridos más cortos y homogéneos. Esto indica una capacidad limitada para diferenciar este tipo de buque frente a otros con trayectorias más locales.

4.3.2 Petroleros

A continuación, se analiza el comportamiento del modelo SeqVAE para el caso de los buques petroleros, cuyo patrón de navegación, al igual que en los cargueros, está asociado a trayectorias de media y larga distancia.

En línea con el análisis realizado en el apartado 4.1, las trayectorias generadas presentan una limitada dispersión espacial, lo que condiciona la capacidad del modelo para reproducir recorridos extensos.

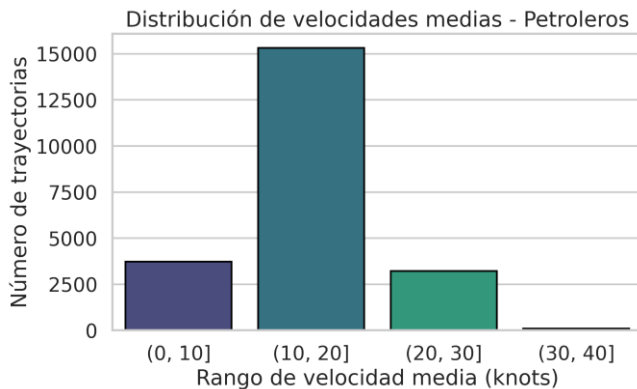


Figura 4-38 Distribución velocidades reales Petroleros

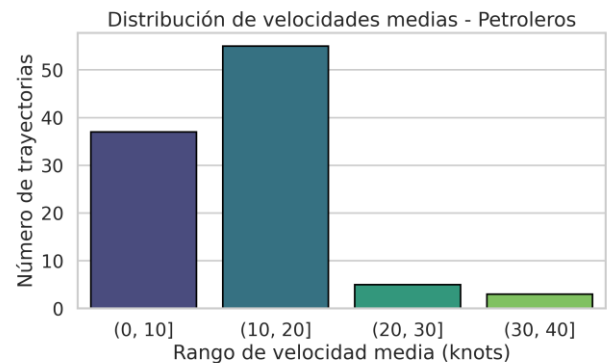


Figura 4-39 Distribución velocidades SeqVAE Petroleros

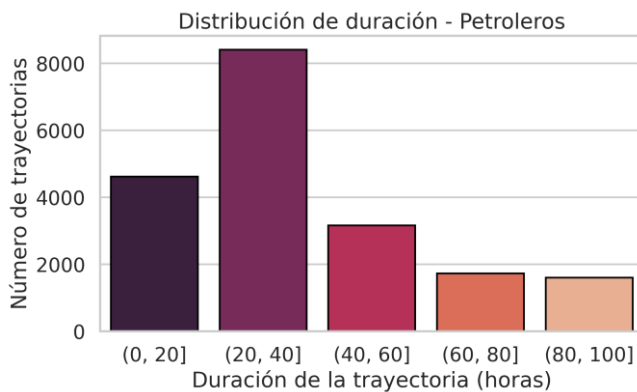


Figura 4-40 Distribución tiempo reales Petroleros

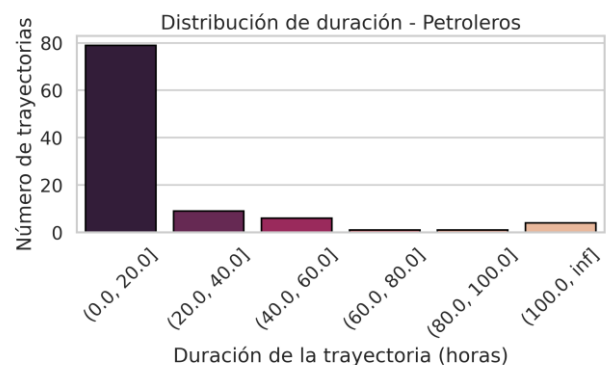


Figura 4-41 Distribución tiempo SeqVAE Petroleros

En las gráficas de **distribución de velocidades, duración y distancia** (Figura 4-38, Figura 4-39, Figura 4-40, Figura 4-41, Figura 4-42 y Figura 4-43) se observa una clara concentración en los intervalos de menor duración y distancia, en contraste con los datos reales, donde la distribución es más amplia y con mayor presencia de trayectorias de duración media. Esto indica que el modelo genera trayectorias más cortas, alejándose del comportamiento esperado para este tipo de buque.

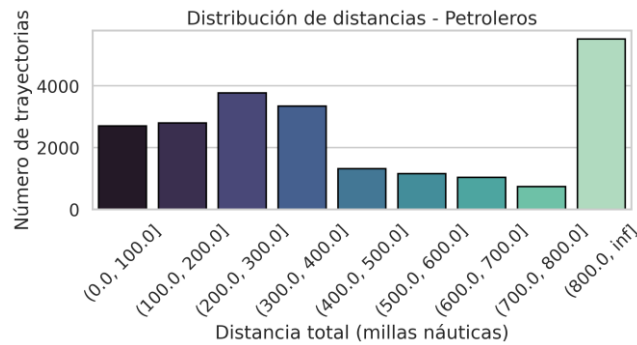


Figura 4-42 Distribución distancias reales Petroleros

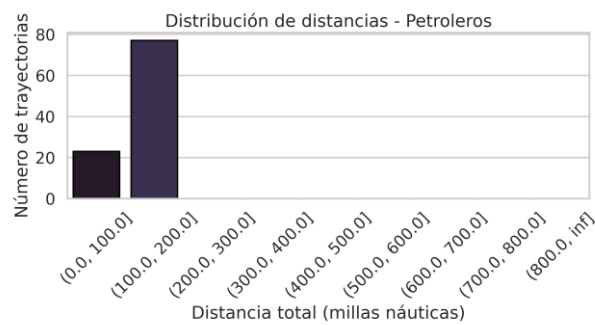


Figura 4-43 Distribución distancias SeqVAE Petroleros

En conjunto, el modelo SeqVAE muestra dificultades para reproducir la escala de las trayectorias propias de los petroleros, generando recorridos más reducidos y homogéneos. Al igual que en el caso de los cargueros, esto refleja una capacidad limitada para capturar trayectorias de largo recorrido y diferenciar adecuadamente este tipo de buque frente a otros con patrones más locales.

4.3.3 Pesqueros

A continuación, se analiza el comportamiento del modelo SeqVAE para el caso de los buques pesqueros, caracterizados por trayectorias cortas y de carácter local.

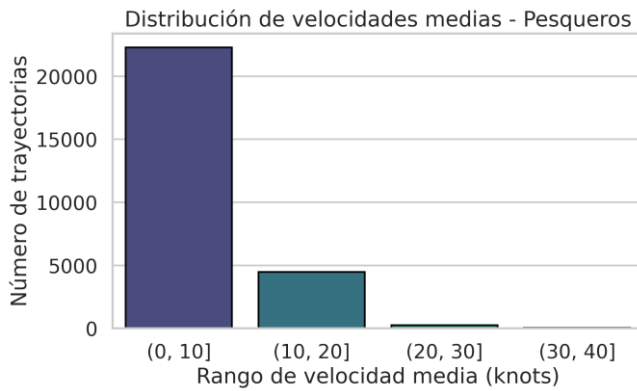


Figura 4-44 Distribución velocidades reales Pesqueros

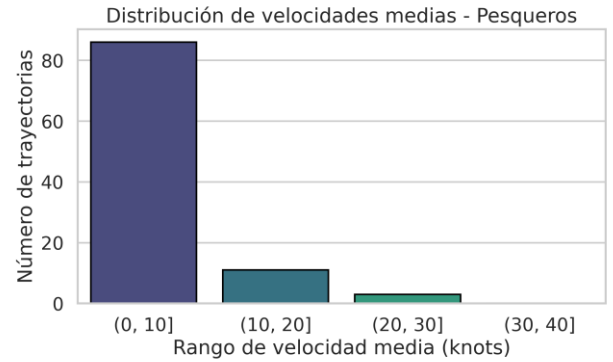


Figura 4-45 Distribución velocidades SeqVAE Pesqueros

En las gráficas de **distribución de velocidades, duración y distancia** (Figura 4-44, Figura 4-45, Figura 4-46, Figura 4-47, Figura 4-48 y Figura 4-49), se aprecia que las trayectorias generadas se concentran mayoritariamente en los intervalos más bajos, en línea con el carácter reducido de este tipo de navegación. No obstante, en comparación con los datos reales, la distribución aparece más concentrada, con una **menor dispersión de valores**, lo que indica que el modelo genera trayectorias más homogéneas.

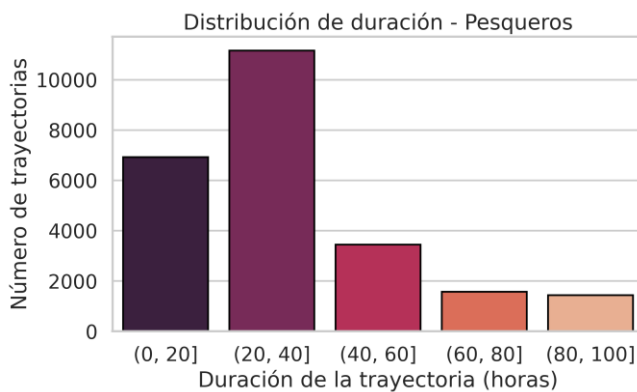


Figura 4-46 Distribución tiempo reales Pesqueros

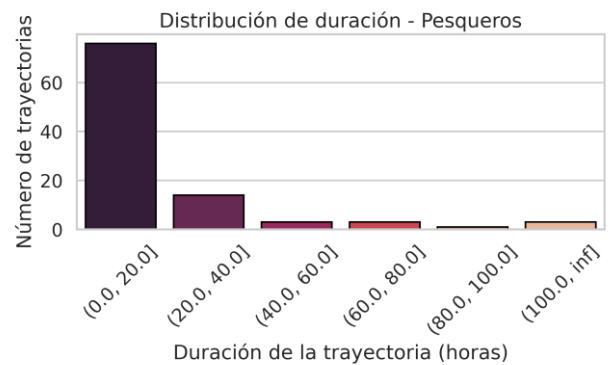


Figura 4-47 Distribución tiempo SeqVAE Pesqueros

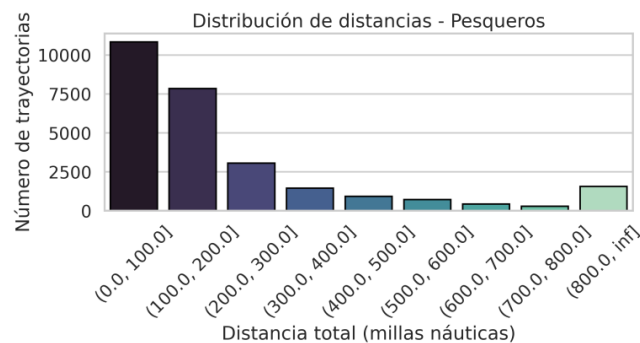


Figura 4-48 Distribución distancias reales Pesqueros

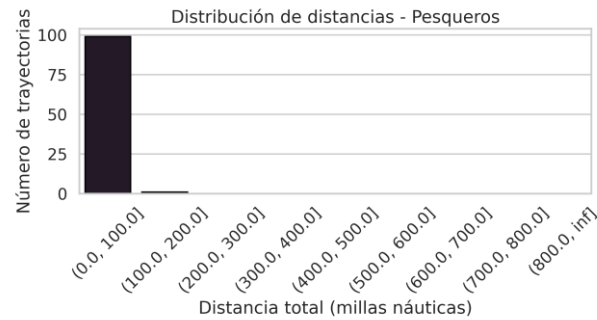


Figura 4-49 Distribución distancias SeqVAE Pesqueros

En conjunto, el modelo SeqVAE reproduce de forma aproximada la escala de las trayectorias de los pesqueros, pero presenta dificultades para capturar su variabilidad espacial y direccional. Esto sugiere que, aunque el modelo se adapta mejor a este tipo de trayectorias cortas que a las de largo recorrido, sigue mostrando limitaciones para representar la diversidad real del comportamiento de estos buques.

4.3.4 Pasajeros

A continuación, se analiza el comportamiento del modelo SeqVAE para el caso de los buques de pasajeros, caracterizados por trayectorias de corta y media distancia, generalmente asociadas a rutas regionales.

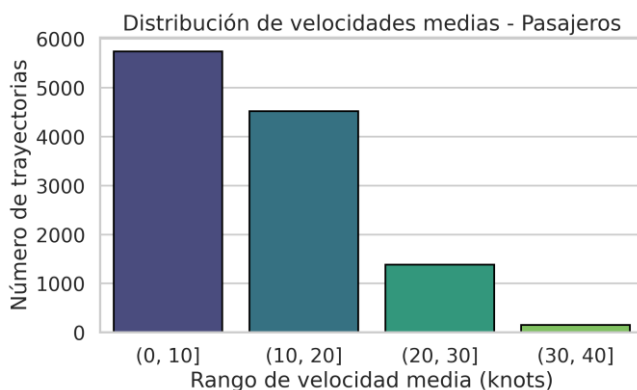


Figura 4-50 Distribución velocidades reales Pasajeros

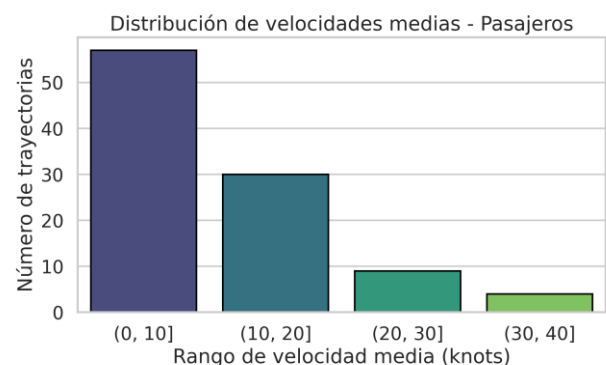


Figura 4-51 Distribución velocidades SeqVAE Pasajeros

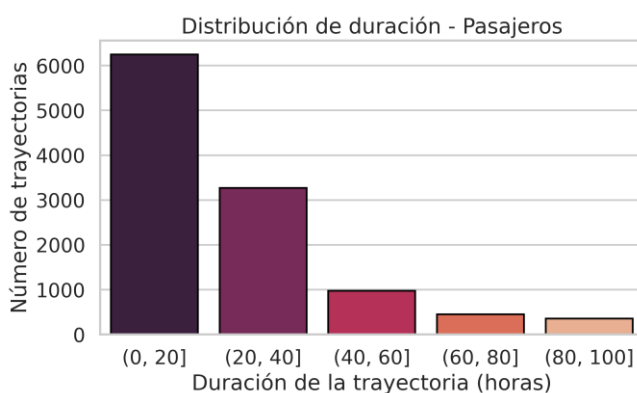


Figura 4-52 Distribución tiempo reales Pasajeros

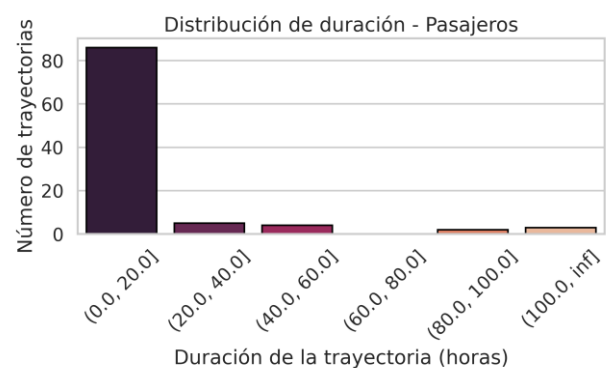


Figura 4-53 Distribución tiempo SeqVAE Pasajeros

En las gráficas de **distribución de velocidades, duración y distancia** (Figura 4-50, Figura 4-51, Figura 4-52, Figura 4-53, Figura 4-54 y Figura 4-55), se observa una concentración en los intervalos de menor duración y distancia, lo que indica que el modelo genera trayectorias más cortas. Aunque este

comportamiento es coherente en escala con este tipo de buque, la distribución resulta más concentrada que en los datos reales, evidenciando una menor diversidad.

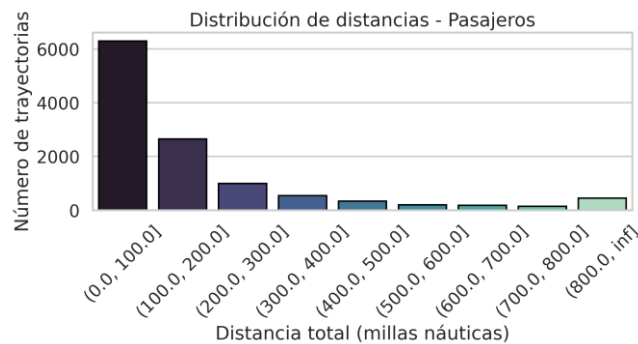


Figura 4-54 Distribución distancia reales Pasajeros

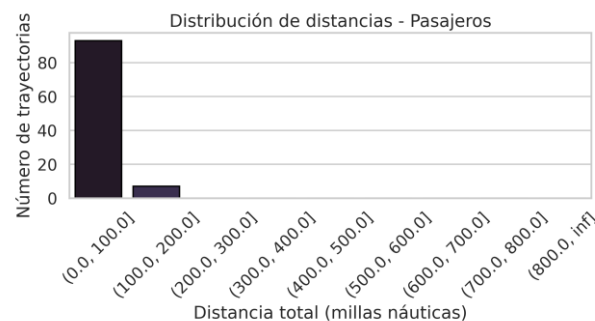


Figura 4-55 Distribución distancia SeqVAE Pasajeros

En conjunto, el modelo SeqVAE reproduce de forma aproximada la escala de las trayectorias de pasajeros, pero presenta una menor capacidad para capturar su variabilidad y dispersión espacial, generando recorridos más homogéneos.

4.3.5 Remolcadores

A continuación, se analiza el comportamiento del modelo SeqVAE para el caso de los remolcadores, caracterizados por trayectorias muy cortas y de carácter local, asociadas a operaciones en entornos portuarios.

En las gráficas de **distribución de velocidades, duración y distancia** (Figura 4-56, Figura 4-57, Figura 4-58, Figura 4-59, Figura 4-60 y Figura 4-61), se observa una clara concentración en los intervalos de menor duración y distancia, en línea con el comportamiento general de este tipo de buque. No obstante, en comparación con los datos reales, la distribución aparece más concentrada, con una presencia muy limitada de trayectorias en intervalos superiores.

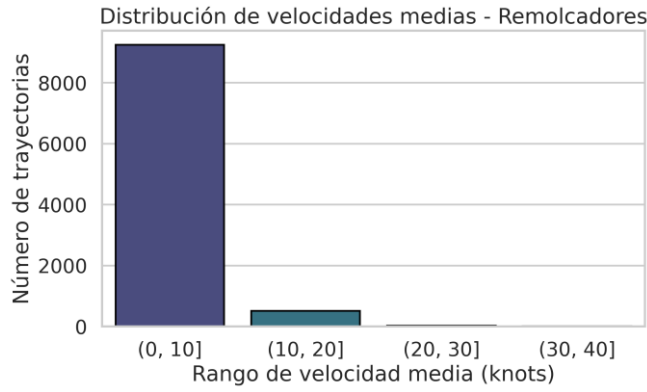


Figura 4-56 Distribución velocidades reales Remolcadores

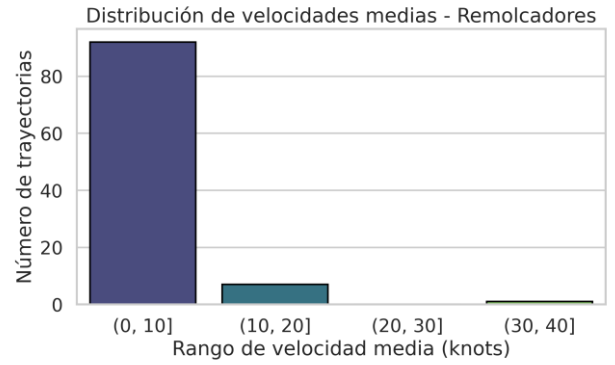


Figura 4-57 Distribución velocidades SeqVAE Remolcadores

Este comportamiento indica que el modelo genera trayectorias más cortas y homogéneas, reproduciendo la escala de las trayectorias, pero sin capturar completamente la variabilidad observada en los datos reales.

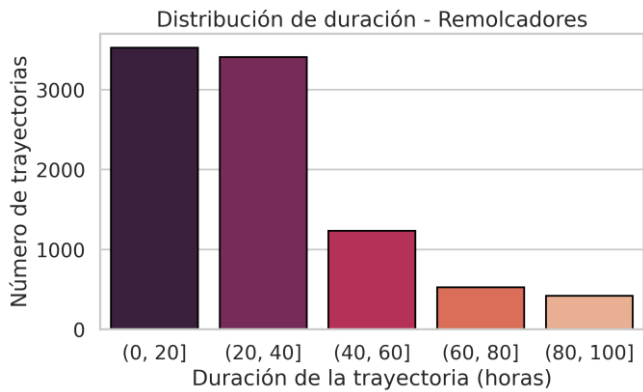


Figura 4-58 Distribución tiempo reales Remolcadores

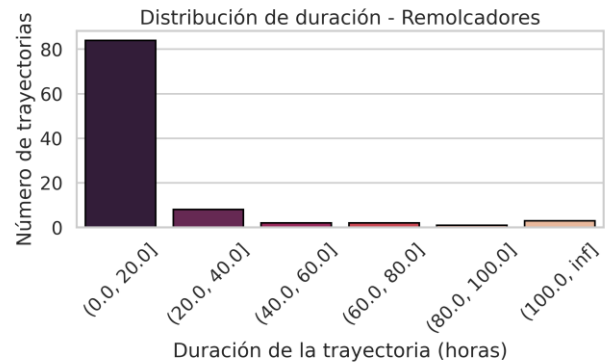


Figura 4-59 Distribución tiempo SeqVAE Remolcadores

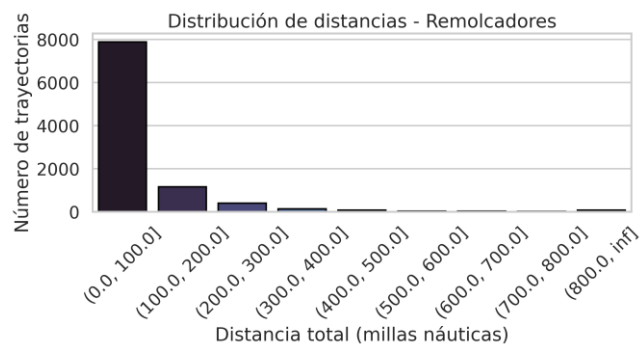


Figura 4-60 Distribución distancia reales Remolcadores

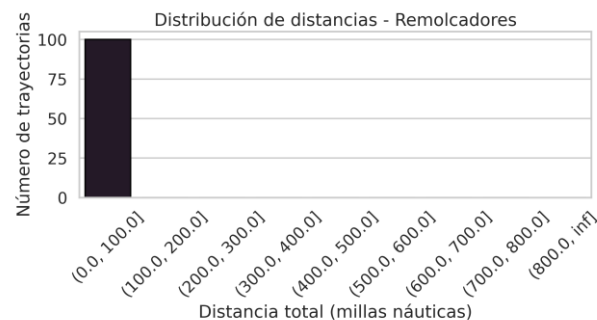


Figura 4-61 Distribución distancias SeqVAE Remolcadores

Considerando todos los resultados, el modelo SeqVAE reproduce de forma aproximada el carácter local de las trayectorias de los remolcadores, pero, al igual que en los casos anteriores, presenta una menor variabilidad y una tendencia a generar recorridos más uniformes, lo que limita su capacidad para diferenciar plenamente este tipo de buque.

4.4 Comparación LSTM y SeqVAE

Los resultados obtenidos permiten establecer diferencias claras entre los modelos LSTM y SeqVAE en la generación de trayectorias sintéticas.

El modelo LSTM muestra una mayor capacidad para reproducir la estructura y diversidad del tráfico marítimo, generando trayectorias que se ajustan de forma razonable a las características de cada tipo de buque. En particular, es capaz de diferenciar entre trayectorias de larga distancia, propias de cargueros y petroleros, y trayectorias más cortas y locales, como las de pesqueros, pasajeros y remolcadores. No obstante, tal y como se ha observado en los apartados anteriores, el modelo presenta una tendencia sistemática a sobreestimar la duración y la distancia de las trayectorias, generando recorridos más largos de lo esperado.

Por otro lado, el modelo SeqVAE presenta un comportamiento más homogéneo, generando trayectorias con menor variabilidad en términos de longitud, duración y forma. Los recorridos generados tienden a ser más cortos en comparación con los datos reales, lo que limita su capacidad para representar trayectorias de mayor extensión, como las de cargueros y petroleros. Esta característica se ha observado de forma consistente en los distintos tipos de buque, donde el modelo tiende a generar patrones más uniformes independientemente de la categoría.

Asimismo, y en línea con lo analizado en el apartado 4.3, la evaluación de la dispersión geográfica en el caso del SeqVAE no resulta representativa, ya que las trayectorias generadas parten de puntos iniciales condicionados por los datos reales. Por este motivo, la comparación entre modelos se centra en características estructurales de las trayectorias, como su longitud, duración y variabilidad.

En conjunto, los resultados indican que el modelo LSTM captura mejor las diferencias estructurales entre tipos de trayectorias, adaptándose a distintos patrones de navegación, aunque con una desviación en la escala de los recorridos. Por su parte, el modelo SeqVAE presenta limitaciones en la representación de la variabilidad y en la diferenciación entre tipos de buque, lo que reduce su capacidad para reproducir de forma realista el tráfico marítimo en su conjunto.

4.5 Incorporación de filtrado de trayectorias sintéticas LSTM

Vista la problemática identificada en el apartado anterior, relacionada con la sobrestimación de la distancia y duración de las trayectorias generadas por el modelo LSTM, se plantean dos estrategias de filtrado alternativas sobre las trayectorias sintéticas con el objetivo de mejorar su coherencia respecto a los datos reales.

La primera estrategia consiste en un **filtrado de control**, basado en la imposición de un límite máximo de distancia total. Este enfoque introduce una restricción fija que permite eliminar aquellas trayectorias cuya longitud resulta claramente irreal o incompatible con el comportamiento observado en los datos reales. Los umbrales definidos, dependientes del tipo de buque, son los siguientes:

- Cargueros: 1000 millas náuticas
- Petroleros: 1200 millas náuticas
- Pesqueros: 500 millas náuticas
- Pasajeros: 500 millas náuticas
- Remolcadores: 200 millas náuticas

En este caso, las trayectorias se generan sin restricciones durante el proceso y se descartan posteriormente si superan dichos límites.

Como alternativa, se propone una segunda estrategia basada en un **filtrado condicionado** por la distribución real de distancias. En este enfoque, cada trayectoria sintética se genera a partir de una distancia objetivo muestreada de los datos reales, de modo que el proceso generativo queda condicionado a reproducir la variabilidad observada. La trayectoria se acepta únicamente si su distancia total se encuentra dentro de un intervalo de tolerancia respecto a dicha distancia objetivo, descartándose en caso contrario. A diferencia del primer método, que introduce una restricción rígida, este segundo enfoque permite capturar de forma más realista la distribución global de distancias, evitando tanto valores extremos como la pérdida de variabilidad.

En la Tabla 4-4, correspondiente a la velocidad media, se observa que los valores obtenidos tras el filtrado se aproximan, en términos generales, a los datos reales. Excepto para el caso de los cargueros con el filtrado de control y el caso de los pasajeros con filtrado condicionado, todos los tipos de buque, las velocidades medias corregidas muestran una mejora respecto a las obtenidas inicialmente con el modelo LSTM, reduciendo las desviaciones observadas. No obstante, aunque el ajuste es significativo, todavía persisten diferencias en algunos casos, lo que indica que no se reproduce completamente la distribución real de este parámetro.

Tipo de buque	Velocidad media reales (nudos)	Velocidad media LSTM (nudos)	Velocidad media con filtrado de control (nudos)	Velocidad media con filtrado condicionado (nudos)
0-Carguero	14.94	12.52	11.74	12.67
1-Petrolero	14.61	13.71	14.29	13.98
2-Pesquero	6.96	9.88	8.62	7.12
3-Pasajero	11.38	12.31	11.02	9.79
4-Remolcador	3.77	8.67	6.70	5.81

Tabla 4-4 Velocidades medias según modelo y tipo de buque

En la Tabla 4-5, relativa a la distancia media, el efecto del filtrado resulta especialmente notable. Las trayectorias generadas por el modelo LSTM sin restricciones presentan una clara sobreestimación de las distancias respecto a los datos reales. Tras la aplicación del filtrado, los valores se reducen de forma considerable, acercándose a los reales en todos los tipos de buque. Esta mejora es particularmente relevante en pesqueros, pasajeros y remolcadores, donde pequeñas desviaciones tienen un mayor impacto debido a la menor escala de sus trayectorias. A pesar de esta corrección, en algunos casos todavía se observan diferencias residuales, lo que refleja limitaciones en el ajuste fino del modelo.

Tipo de buque	Distancia media reales (millas)	Distancia media LSTM (millas)	Distancia media con filtrado de control (nudos)	Distancia media con filtrado condicionado (nudos)
0-Carguero	687.52	975.15	881.73	436.08
1-Petrolero	637.36	1129.08	616.03	442.36
2-Pesquero	252.02	1073.40	348.14	179.69
3-Pasajero	194.09	550.43	314.59	159.09
4-Remolcador	79.67	550.43	271.85	50.51

Tabla 4-5 Distancias medias según modelo y tipo de buque

Por su parte, en la Tabla 4-6, correspondiente a la duración media, se observa un comportamiento coherente con el de la distancia. La reducción en la longitud de las trayectorias conlleva una disminución en la duración media, permitiendo un mejor ajuste a los valores reales. Las diferencias se reducen de manera significativa en todos los tipos de buque, especialmente cuando se aplica el filtrado condicionado. Sin embargo, en algunos casos (como los cargueros y petroleros) el ajuste puede resultar excesivo, lo que sugiere que el filtrado introduce cierto sesgo hacia la reducción de las trayectorias.

Tipo de buque	Duración media reales (horas)	Duración media LSTM (horas)	Duración media con filtrado de control (horas)	Duración media con filtrado condicionado (horas)
0-Carguero	58.87	443.30	75.10	63.51
1-Petrolero	53.13	236.55	43.11	41.64
2-Pesquero	47.27	134.92	40.39	35.23
3-Pasajero	30.36	55.72	28.55	36.25
4-Remolcador	40.11	72.58	40.57	28.64

Tabla 4-6 Duración media según modelo y tipo de buque

Por otro lado, la Tabla 4-7 recoge los tiempos de generación de la Tabla 3-3 para el modelo LSTM añadiendo los tiempos de generación del modelo bajo las distintas configuraciones. Se observa que la aplicación del filtrado de control introduce un incremento significativo en el tiempo de generación respecto al modelo original, mientras que el filtrado condicionado supone una disminución considerable del coste computacional. Este comportamiento se explica por la necesidad de aplicar restricciones adicionales durante el proceso de generación, lo que incrementa el número de iteraciones necesarias para obtener trayectorias válidas.

Tipo de buque	Tiempo de generación LSTM original	Tiempo de generación LSTM con filtrado de control	Tiempo de generación LSTM con filtrado condicionado
0-Carguero	43m 18s	39m 31.03s	3m 40.23s
1-Petrolero	16m 54.13s	22m 49.32s	6m 29.96s
2-Pesquero	32m 44.61s	1h 16m 19.18s	4m 49s
3-Pasajero	29m 52,8s	1h 16m 58.89s	4m 56.36s
4-Remolcador	1h 11m 47.6s	1h 32m 59.71s	10m 16.36s

Tabla 4-7 Tiempos de generación con variantes del modelo LSTM según modelo y tipo de buque

Los resultados obtenidos muestran que el filtrado aplicado permite corregir de forma efectiva la sobreestimación de distancia y duración presente en el modelo LSTM, mejorando el ajuste de las trayectorias sintéticas a los datos reales. Sin embargo, esta mejora no está exenta de compromisos, ya que en algunos casos se introduce una ligera pérdida de variabilidad. Esto pone de manifiesto la necesidad de equilibrar precisión y eficiencia en el diseño del modelo.

4.5.1 Cargueros

A continuación, se analiza el comportamiento del modelo LSTM tras la aplicación del filtrado condicionado para todos los tipos de buques, empezando por los cargueros, a partir de las distribuciones obtenidas.

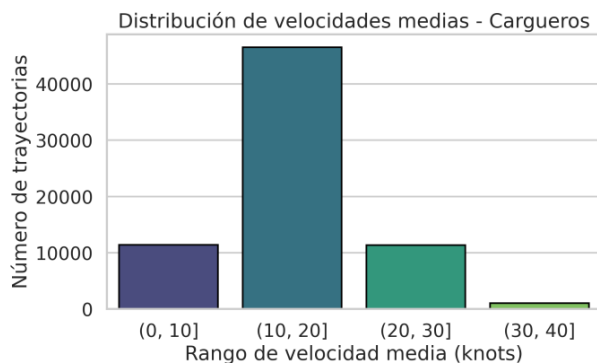


Figura 4-62 Distribución velocidades reales Cargueros

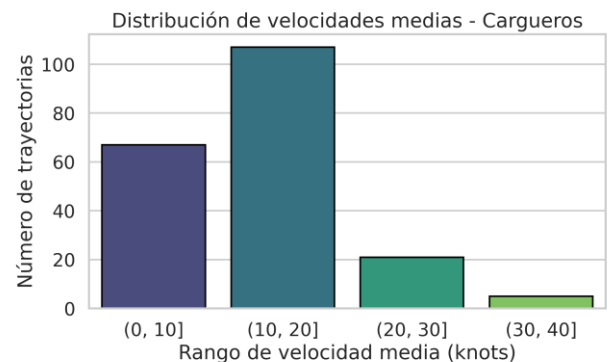


Figura 4-63 Distribución velocidades LSTM con filtrado condicionado para Cargueros

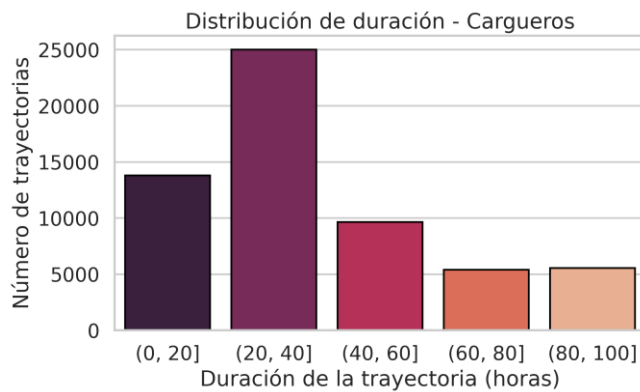


Figura 4-64 Distribución tiempo reales Cargueros

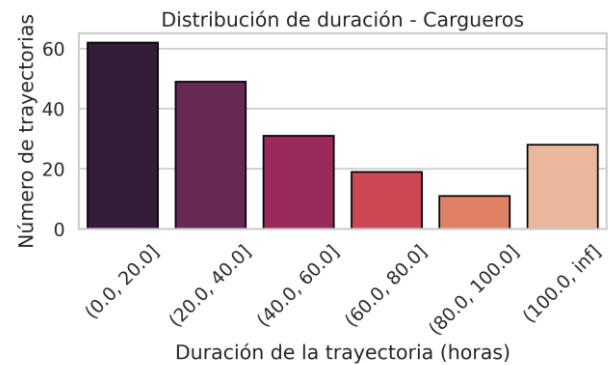


Figura 4-65 Distribución tiempo LSTM con filtrado condicionado para Cargueros

En las **gráficas de distribución de velocidades medias** (Figura 4-62 y Figura 4-63) se observa que las trayectorias generadas mantienen la forma general de los datos reales, con una clara concentración en el intervalo de 10 a 20 nudos. El modelo reproduce correctamente el rango predominante de velocidades, lo que indica que esta variable está bien capturada.

En cuanto a la **distribución de la duración de las trayectorias** (Figura 4-64 y Figura 4-65), se observa un cambio más significativo. Las trayectorias generadas presentan una mayor concentración en intervalos de duración intermedios, reduciéndose notablemente la presencia de trayectorias muy largas. Aunque la forma general de la distribución se mantiene, este comportamiento confirma que el filtrado tiende a acortar las trayectorias, lo que mejora el ajuste global, pero introduce una desviación respecto a los valores más altos observados en los datos reales.

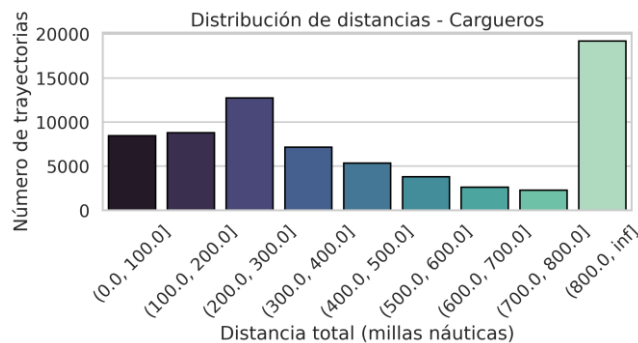


Figura 4-66 Distribución distancias reales Cargueros

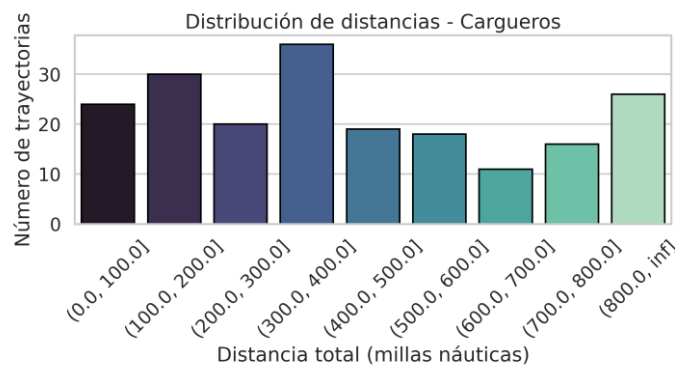


Figura 4-67 Distribución distancias LSTM con filtrado condicionado para Cargueros

La **distribución de distancias** (Figura 4-66 y Figura 4-67) muestra de forma más clara el efecto del filtrado. Mientras que los datos reales presentan una distribución más equilibrada entre intervalos, las

trayectorias generadas concentran una mayor proporción en rangos de distancia media, reduciendo significativamente la presencia de valores extremos. Esto indica que el modelo, tras el filtrado, limita la generación de trayectorias excesivamente largas, aunque sigue existiendo una ligera tendencia a sobreestimar ciertos rangos intermedios.

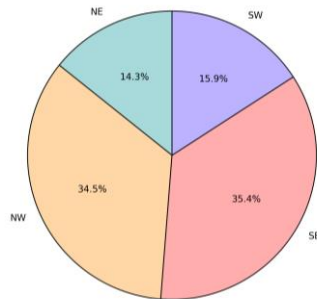


Figura 4-68 Distribución geográfica reales Cargueros

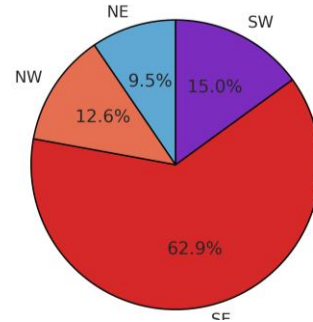


Figura 4-69 Distribución geográfica LSTM con filtrado condicionado para Cargueros

Por último, en la **distribución geográfica por cuadrantes** (Figura 4-68 y Figura 4-69), se observan diferencias más marcadas entre los datos reales y los generados. Mientras que las trayectorias reales presentan una distribución más equilibrada entre cuadrantes, las trayectorias sintéticas muestran una clara concentración en el cuadrante SE, con una reducción notable en otros cuadrantes. Este comportamiento evidencia una pérdida de diversidad geográfica, lo que sugiere que el modelo presenta limitaciones para reproducir correctamente la dispersión espacial de las trayectorias.

En términos globales, el modelo LSTM tras la aplicación del filtrado mantiene la capacidad de reproducir las características principales de las trayectorias, especialmente en variables como la velocidad. Sin embargo, persisten limitaciones menores en la representación de la duración, la distancia y, especialmente, la distribución geográfica.

4.5.2 Petroleros

Tras la aplicación del filtrado sobre el método de generación del modelo LSTM, se analizan las distribuciones resultantes para el caso de los buques petroleros, comparándolas con los datos reales.

En las gráficas de **distribución de velocidades medias** (Figura 4-70 y Figura 4-71), el modelo reproduce correctamente el intervalo dominante de 10 a 20 nudos, sin que el filtrado introduzca cambios relevantes. La estructura general se mantiene, aunque se observa una ligera reducción en los valores extremos, indicando una menor dispersión.

En cuanto a la **distribución de la duración** de las trayectorias (Figura 4-72 y Figura 4-73), muestra un desplazamiento hacia intervalos intermedios, con una reducción clara de trayectorias de larga duración. Este ajuste mejora la coherencia global, pero limita la representación de los casos más extensos.

En cuanto a la **distribución de distancias** (Figura 4-74 y Figura 4-75), se confirma esta tendencia: disminuye la presencia de trayectorias excesivamente largas y aumenta la concentración en rangos medios. El filtrado corrige la sobreestimación inicial, aunque introduce cierta compresión en la escala de los recorridos.

Por último, en la **distribución geográfica por cuadrantes** (Figura 4-76 y Figura 4-77), evidencia una pérdida de equilibrio espacial. Frente a la distribución más uniforme de los datos reales, las trayectorias generadas se concentran principalmente en los cuadrantes SE y NW, reduciendo la diversidad geográfica.

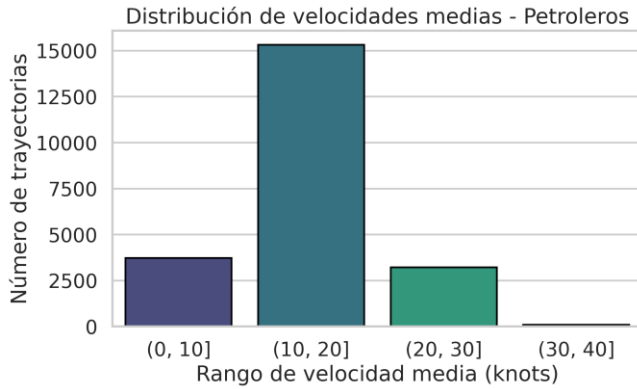


Figura 4-70 Distribución velocidades reales Petroleros

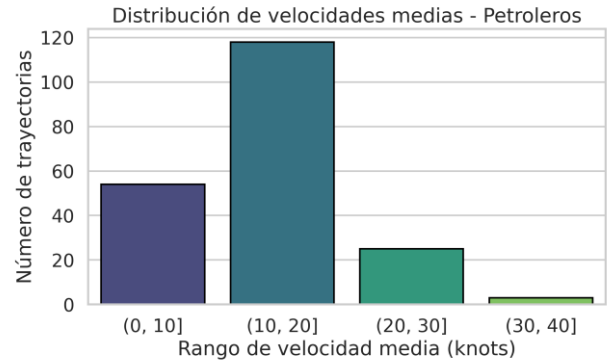


Figura 4-71 Distribución velocidades LSTM con filtrado condicionado para Petroleros

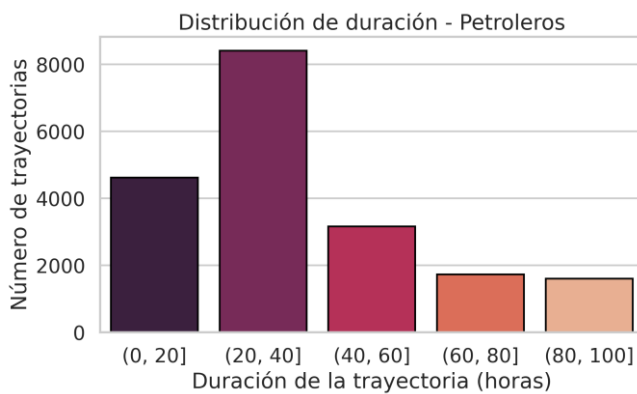


Figura 4-72 Distribución tiempo reales Petroleros

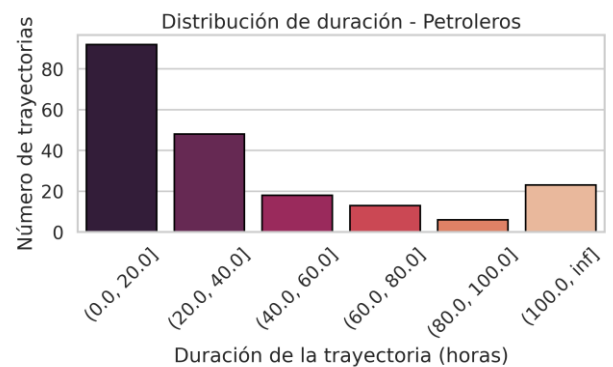


Figura 4-73 Distribución tiempo LSTM con filtrado condicionado para Petroleros

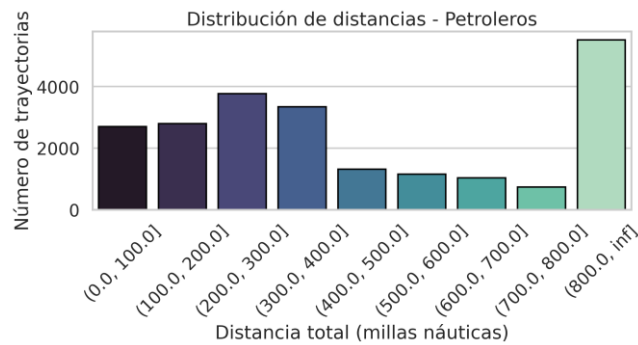


Figura 4-74 Distribución distancias reales Petroleros

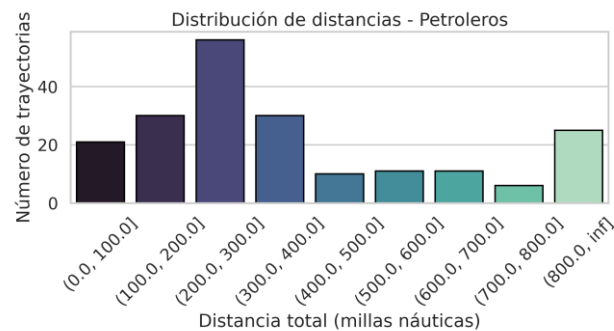


Figura 4-75 Distribución distancias LSTM con filtrado condicionado para Petroleros

En conjunto, el modelo mantiene la estructura general del comportamiento de los petroleros, pero el filtrado, si bien corrige la escala de las trayectorias, reduce la variabilidad y acentúa ciertas concentraciones espaciales, reflejando limitaciones en la reproducción completa de la distribución real.

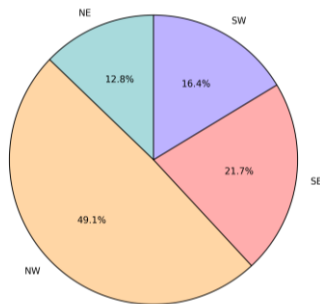


Figura 4-76 Distribución geográfica reales Petroleros

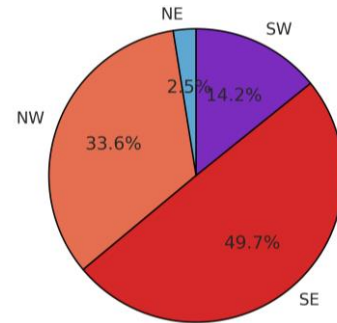


Figura 4-77 Distribución geográfica LSTM con filtrado condicionado para Petroleros

4.5.3 Pesqueros

A continuación, se analiza el comportamiento del modelo LSTM tras el filtrado para el caso de los buques pesqueros, destacando los aspectos más característicos de este tipo de navegación.

En las **gráficas de velocidad** (Figura 4-78 y Figura 4-79) el modelo reproduce adecuadamente el patrón dominante de bajas velocidades, con una clara concentración en el intervalo de 0 a 10 nudos. El filtrado no introduce cambios relevantes, aunque se aprecia una ligera reducción en la dispersión, especialmente en los intervalos menos representativos.

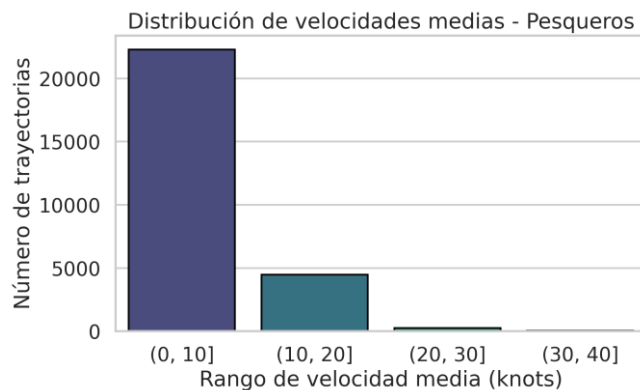


Figura 4-78 Distribución velocidades reales Pesqueros

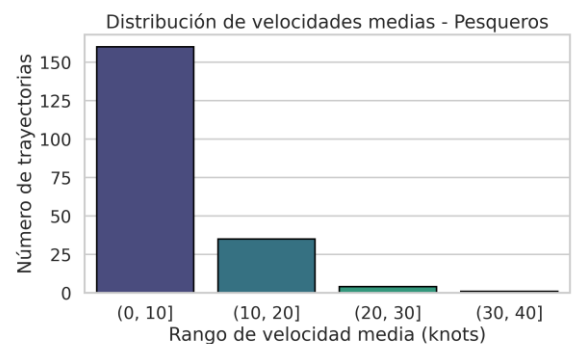


Figura 4-79 Distribución velocidades LSTM con filtrado condicionado para Pesqueros

En cuanto a la **duración** (Figura 4-80 y Figura 4-81), muestra un desplazamiento hacia valores intermedios, reduciéndose la presencia de trayectorias de mayor duración. Este comportamiento mejora el ajuste global, pero atenúa la representación de trayectorias más largas, que son relevantes en determinados contextos operativos.

Este efecto se refleja de forma más evidente en la **distribución de distancias** (Figura 4-82 y Figura 4-83), se confirma esta tendencia: disminuyen los recorridos más extensos y aumenta la concentración en rangos cortos y medios. El filtrado corrige la sobreestimación inicial del modelo, aunque introduce una cierta compresión en la escala de las trayectorias, especialmente significativa en este tipo de buque.

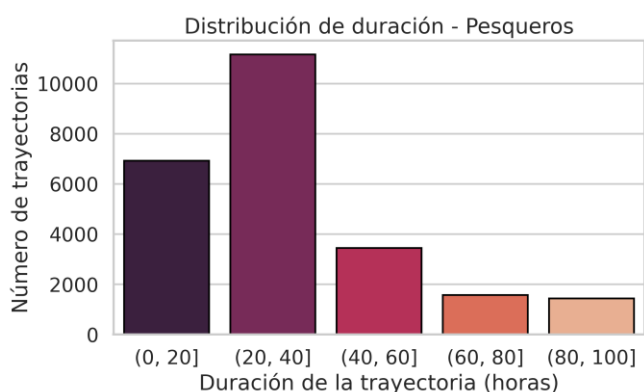


Figura 4-80 Distribución tiempo reales Pesqueros

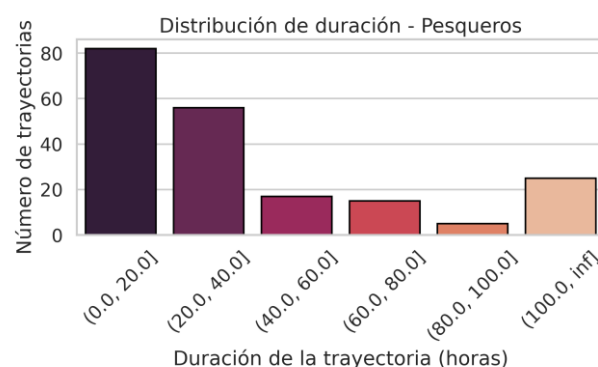


Figura 4-81 Distribución tiempo LSTM con filtrado condicionado para Pesqueros

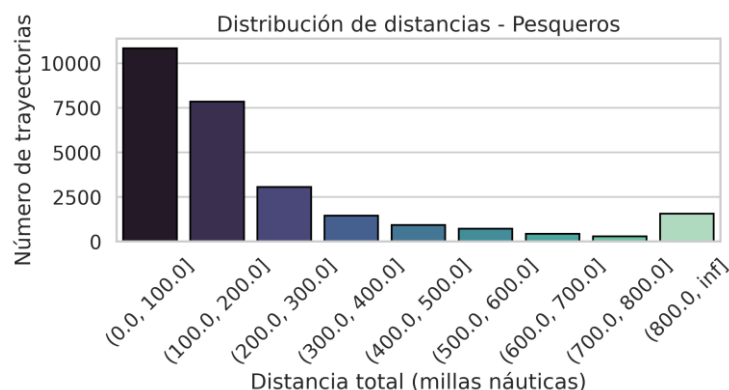


Figura 4-82 Distribución distancias reales Pesqueros

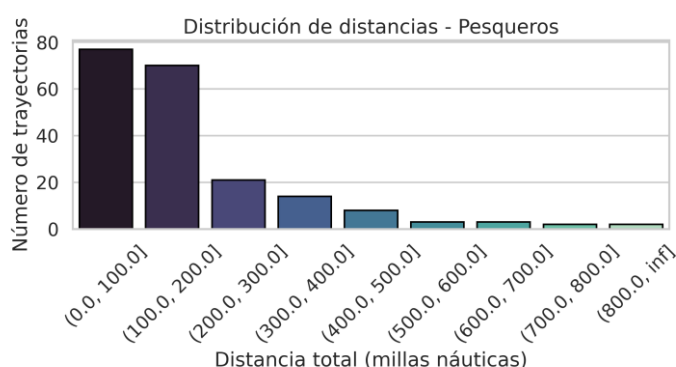


Figura 4-83 Distribución distancias LSTM con filtrado condicionado para Pesqueros

Por último, en la **distribución geográfica** (Figura 4-84 y Figura 4-85), muestra una pérdida de equilibrio espacial. Mientras que los datos reales presentan una dispersión más uniforme con una concentración ligeramente superior en el cuadrante SW, las trayectorias generadas se concentran principalmente en los cuadrantes SE y NW, reduciendo la variabilidad direccional y alejándose del carácter más local y disperso de los pesqueros.

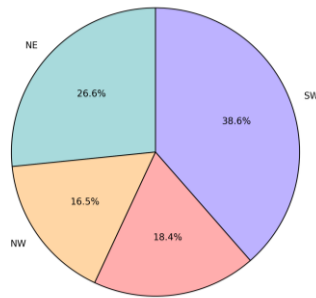


Figura 4-84 Distribución geográfica reales Pesqueros

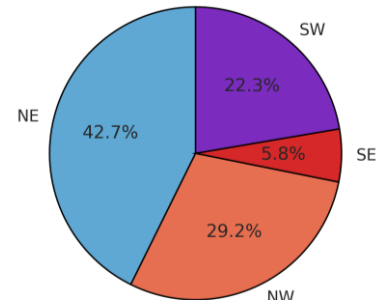


Figura 4-85 Distribución geográfica LSTM con filtrado condicionado para Pesqueros

El comportamiento observado indica que, aunque el modelo captura correctamente el régimen de baja velocidad característico de estos buques, presenta limitaciones en la representación de la extensión y orientación de las trayectorias, generando recorridos más homogéneos y menos diversos que los reales.

4.5.4 Pasajeros

A continuación, se analiza el comportamiento del modelo LSTM tras el filtrado para el caso de los buques de pasajeros, caracterizados por trayectorias de corta y media distancia.

En las **gráficas de velocidad** (Figura 4-86 y Figura 4-87) se mantiene la estructura característica de este tipo de buque, con una concentración principal en los intervalos de 0 a 20 nudos. El filtrado no introduce cambios relevantes, aunque se observa una ligera reducción en la presencia de velocidades más altas, lo que indica una menor dispersión en esta variable.

En las **distribuciones de duración y distancia** (Figura 4-88, Figura 4-89, Figura 4-90 y Figura 4-91) se aprecia una concentración más acusada en los intervalos bajos, con una disminución progresiva en los valores medios y altos. Las trayectorias generadas presentan menor presencia en los rangos más extensos que en los rangos más cortos, lo que sugiere una cierta compresión en la escala de los recorridos, también presente en los datos reales.

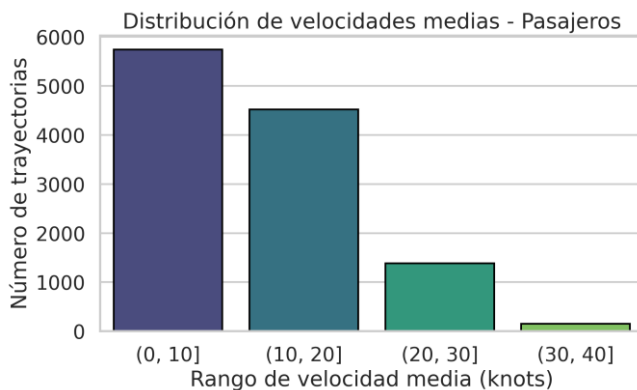


Figura 4-86 Distribución velocidades reales Pasajeros

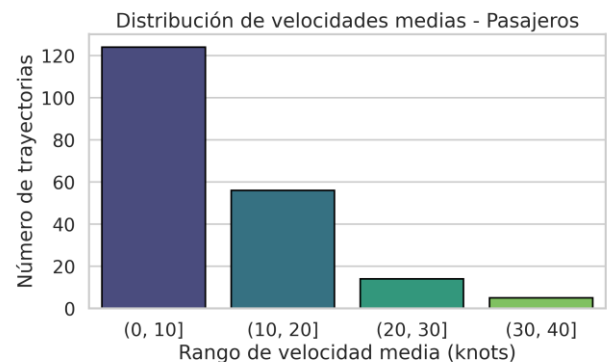


Figura 4-87 Distribución velocidades LSTM con filtrado condicionado para Pasajeros

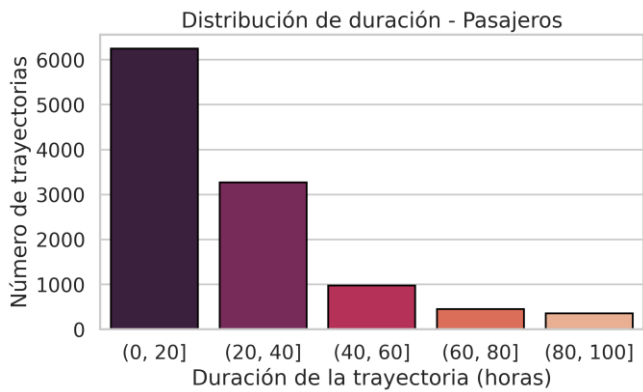


Figura 4-88 Distribución tiempo reales Pasajeros

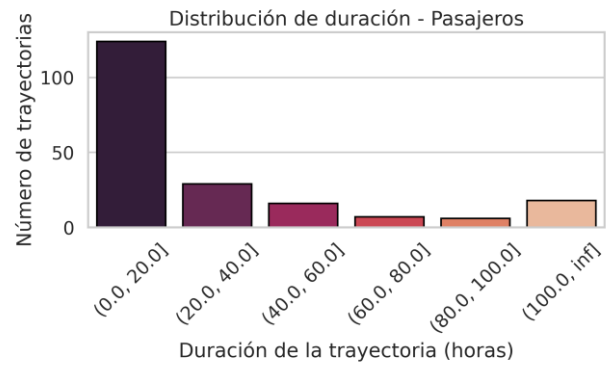


Figura 4-89 Distribución tiempo LSTM con filtrado condicionado para Pasajeros

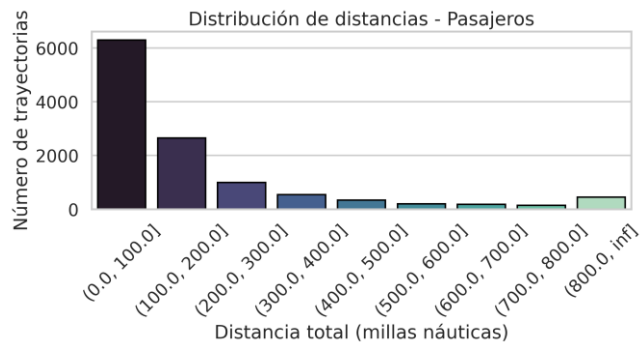


Figura 4-90 Distribución distancia reales Pasajeros

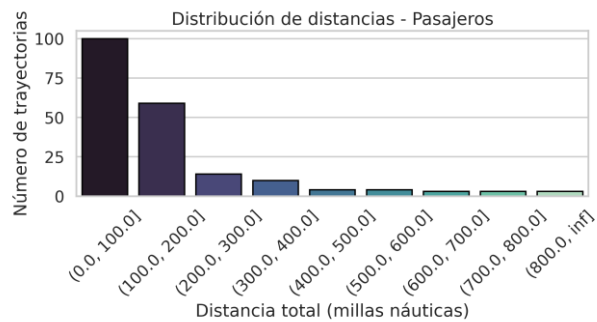


Figura 4-91 Distribución distancia LSTM con filtrado condicionado para Pasajeros

En cuanto a la **distribución geográfica** (Figura 4-92 y Figura 4-93), se observan diferencias significativas. Mientras que los datos reales presentan una distribución más equilibrada con cierta predominancia en el cuadrante NW (42.3%), las trayectorias generadas se concentran principalmente en el cuadrante SE, con presencia secundaria en SW y NW, y ausencia prácticamente total (0.2%) en el cuadrante NE. Esta falta de representación en una dirección concreta evidencia una pérdida de diversidad direccional.

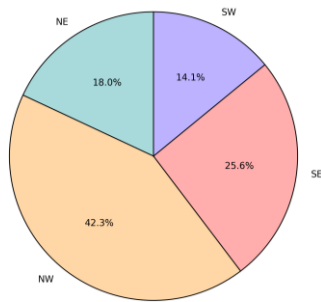


Figura 4-92 Distribución geográfica reales Pasajeros

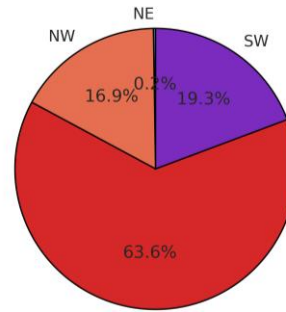


Figura 4-93 Distribución geográfica LSTM con filtrado condicionado para Pasajeros

El análisis indica que, aunque el modelo reproduce adecuadamente el rango de velocidades característico de los buques de pasajeros, el filtrado introduce una reducción de la variabilidad tanto en la escala de las trayectorias como en su dispersión espacial, generando recorridos más concentrados y menos representativos del comportamiento real.

4.5.5 Remolcadores

A continuación, se analiza el comportamiento del modelo LSTM tras el filtrado para el caso de los remolcadores, caracterizados por trayectorias muy cortas y de ámbito local.

En las **gráficas de velocidad** (Figura 4-94 y Figura 4-95) se mantiene de forma clara la concentración en el intervalo de 0 a 10 nudos, reproduciendo correctamente el patrón de navegación a baja velocidad propio de este tipo de buque. El filtrado no introduce cambios apreciables, lo que indica que esta variable se conserva adecuadamente.

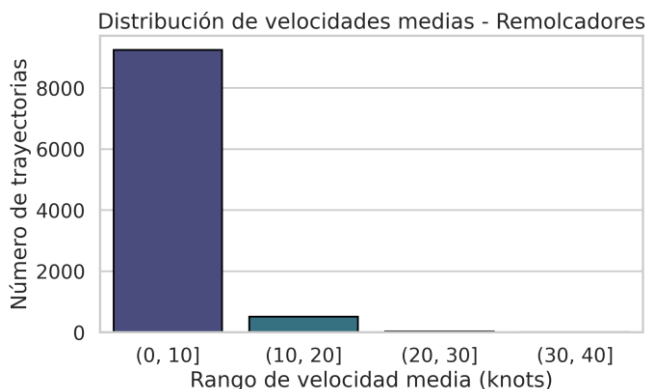


Figura 4-94 Distribución velocidades reales Remolcadores

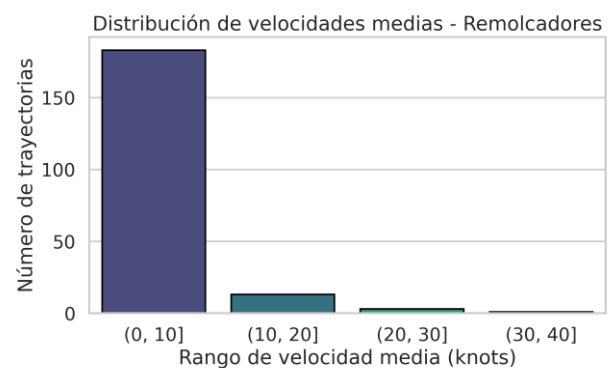


Figura 4-95 Distribución velocidades LSTM con filtrado condicionado para Remolcadores

Sin embargo, en las **distribuciones de duración y distancia** (Figura 4-96, Figura 4-97, Figura 4-98 y Figura 4-99) se observa un efecto distinto al de otros tipos de buque. Las trayectorias generadas se concentran de forma muy marcada en los intervalos más bajos, reduciendo significativamente la presencia de valores intermedios y altos. Este comportamiento implica una simplificación de la distribución, con una pérdida de variabilidad y una menor representación de trayectorias algo más largas que sí aparecen en los datos reales.

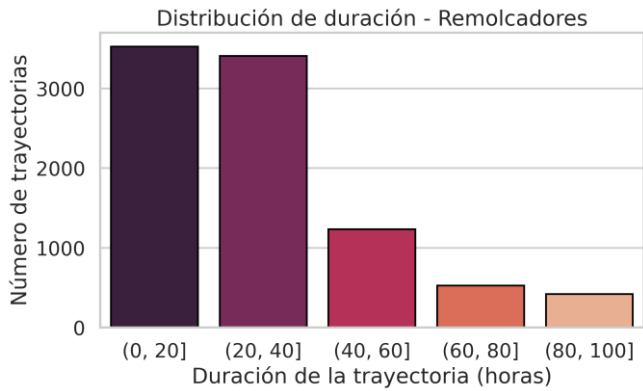


Figura 4-96 Distribución tiempo reales Remolcadores

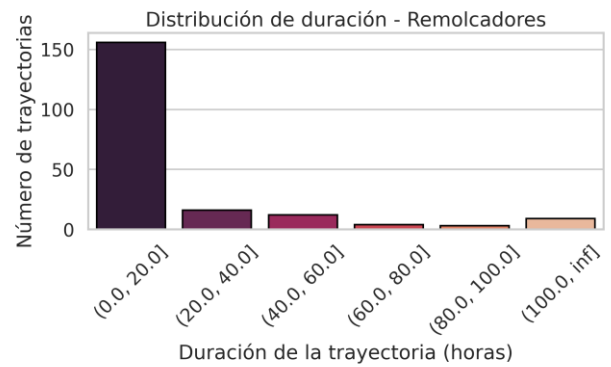


Figura 4-97 Distribución tiempo LSTM con filtrado condicionado para Remolcadores

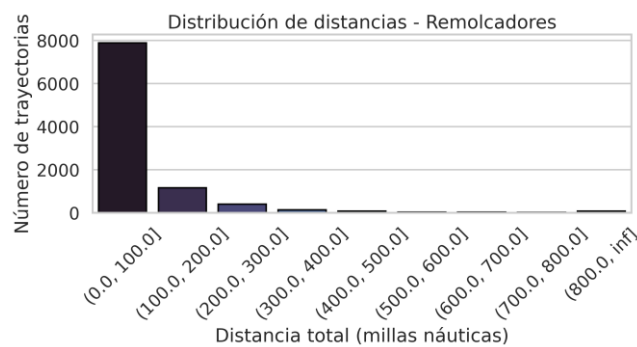


Figura 4-98 Distribución distancia reales Remolcadores

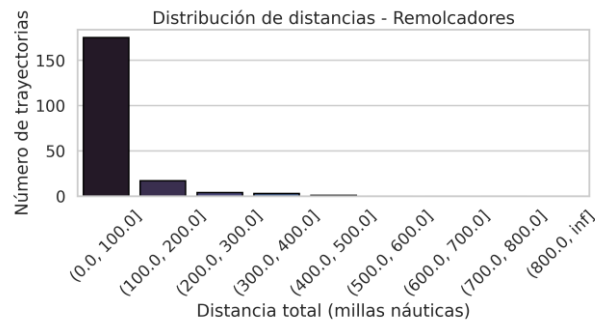


Figura 4-99 Distribución distancias LSTM con filtrado condicionado para Remolcadores

En cuanto a la **distribución geográfica** (Figura 4-100 y Figura 4-101), se aprecia una fuerte concentración en los cuadrantes SE y NW, esta concentración se puede apreciar también en los datos reales. De esta manera, se observa que el modelo es capaz de apreciar la mayor presencia de este tipo de buque en las zonas geográficas mencionadas.

El análisis muestra que, aunque el modelo reproduce correctamente el régimen de baja velocidad característico de los remolcadores, el filtrado tiende a concentrar en exceso las trayectorias en rangos muy reducidos y en determinadas direcciones, limitando la variabilidad tanto en escala como en dispersión espacial.

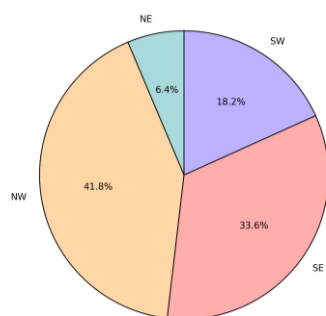


Figura 4-100 Distribución geográfica reales Remolcadores

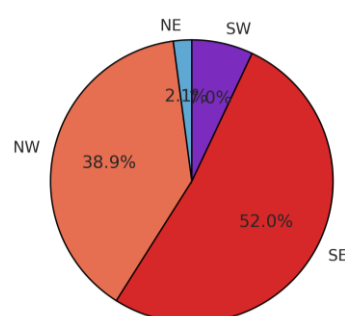


Figura 4-101 Distribución geográfica LSTM con filtrado condicionado para Remolcadores

En conjunto, el modelo LSTM, tras la aplicación del filtrado, muestra una mejora significativa en la representación de las trayectorias sintéticas, especialmente en lo relativo al control de la escala de los recorridos. El modelo mantiene su capacidad para reproducir los patrones característicos de cada tipo de buque, diferenciando entre trayectorias de largo recorrido y trayectorias de carácter local. No obstante, persisten ciertas limitaciones en la precisión de la duración, la distancia y la dispersión geográfica, lo que indica que, aunque el filtrado corrige parte de las desviaciones iniciales, el modelo aún presenta dificultades para ajustar completamente la variabilidad observada en los datos reales.

4.5.6 Inspección visual de las trayectorias LSTM Filtradas

Una vez aplicadas los filtros al modelo LSTM, se realiza una inspección visual de las trayectorias generadas con el objetivo de evaluar su coherencia espacial y compararlas con los patrones observados en los datos reales (Figura 4-102, Figura 4-103, Figura 4-104, Figura 4-105, Figura 4-106, Figura 4-107 y Figura 4-108).

Siguiendo el criterio establecido en el apartado 4.1, se representan 50 trayectorias sintéticas por tipo de buque, lo que permite analizar los resultados evitando la sobrecarga visual. Esta representación facilita la identificación de patrones generales y posibles desviaciones del modelo.

En conjunto, el modelo LSTM muestra una evolución clara al comparar sus distintas versiones. Frente al modelo inicial (Anexo III), donde se observaban trayectorias erráticas, con elevada dispersión y falta de coherencia espacial, el segundo filtrado (Anexo IV) consigue generar trayectorias visualmente más consistentes, con recorridos más suaves y alineados con patrones de navegación plausibles. Tal y como se aprecia en las figuras, especialmente en las ampliaciones para pasajeros y remolcadores, desaparecen en gran medida las trayectorias anómalas y se obtiene un conjunto más limpio y estructurado.

Sin embargo, esta mejora viene acompañada de una reducción notable en la diversidad. En las imágenes del segundo filtrado se observa una concentración clara de trayectorias en determinadas regiones geográficas (particularmente en zonas del hemisferio sur y en direcciones dominantes), con menor cobertura global en comparación con los datos de trayectoria reales (apartado 4.1). Esta limitación es especialmente evidente en casos como pasajeros y remolcadores, donde las trayectorias tienden a agruparse y perder variabilidad direccional.

Por tanto, el modelo LSTM, tras el segundo filtrado, logra un compromiso entre realismo y control de errores, mejorando significativamente la calidad individual de las trayectorias respecto a la versión inicial. No obstante, esta mejora se produce a costa de una menor capacidad para reproducir la diversidad espacial del tráfico marítimo real. Esto refuerza la idea de que, aunque el enfoque es válido, resulta necesario incorporar mecanismos adicionales que permitan preservar la distribución geográfica y la diversidad de trayectorias sin penalizar el realismo.



Figura 4-102 Representación de 50 trayectorias sintéticas LSTM con filtrado condicionado para Cargueros



Figura 4-103 Representación de 50 trayectorias sintéticas LSTM con filtrado condicionado para Petroleros



Figura 4-104 Representación de 50 trayectorias sintéticas LSTM con filtrado condicionado para Pesqueros

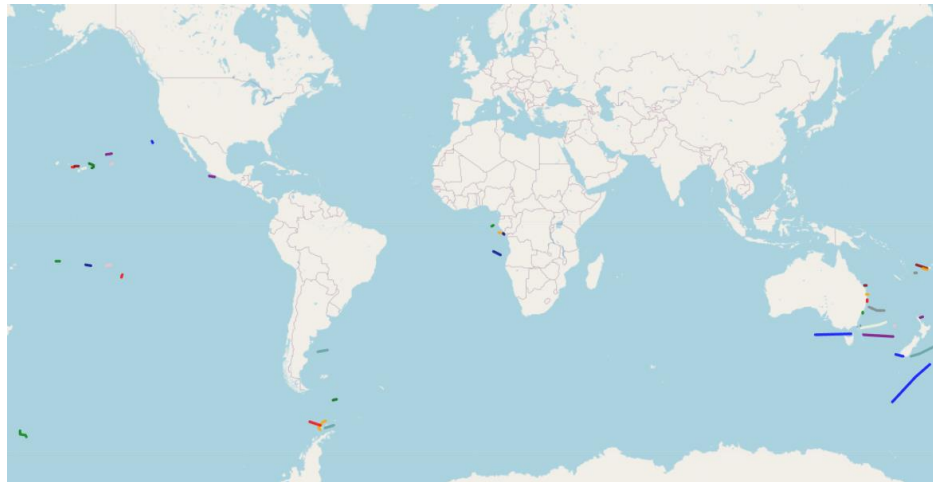


Figura 4-105 Representación de 50 trayectorias sintéticas LSTM con filtrado condicionado para Pasajeros

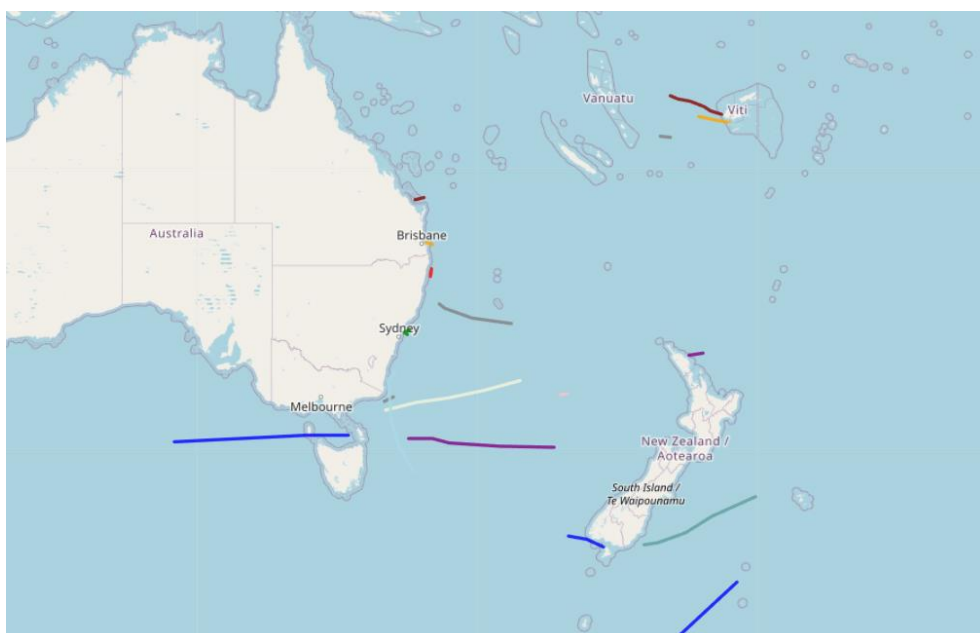


Figura 4-106 Representación ampliada de trayectorias sintéticas LSTM con filtrado condicionado para Pasajeros



Figura 4-107 Representación de 50 trayectorias sintéticas LSTM con filtrado condicionado para Remolcadores

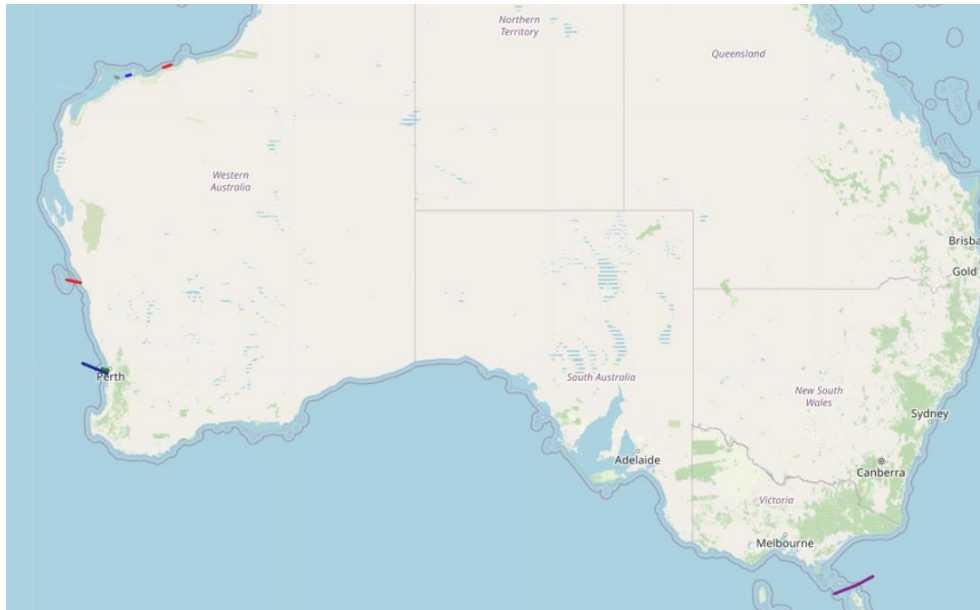


Figura 4-108 Representación ampliada de trayectorias sintéticas LSTM con filtrado condicionado para Remolcadores

4.6 Comparación final de los modelos según el tipo de buque

A continuación, se presenta la comparación final de los modelos considerados óptimos para la generación de trayectorias sintéticas en función del tipo de buque analizado. Tras el desarrollo y evaluación de los distintos modelos propuestos, se ha observado que el rendimiento de cada modelo no es homogéneo, sino que depende de las características específicas del tráfico asociado a cada categoría de buque.

Con el objetivo de sintetizar estos resultados, en la Tabla 4-8 se presenta una visión global en la que se identifica, para cada tipo de buque, el modelo que ofrece el mejor equilibrio entre realismo cinemático y coherencia estadística. Esta comparación permite no solo establecer una selección óptima de modelos, sino también evidenciar la influencia de la dinámica operativa de cada tipo de embarcación en el comportamiento de los modelos generativos.

De este modo, el análisis realizado proporciona una base sólida para la aplicación práctica de los modelos desarrollados, facilitando la elección del enfoque más adecuado en función del contexto de uso.

De manera general, los resultados obtenidos permiten concluir que no existe un único modelo universalmente óptimo para la generación de trayectorias sintéticas, sino que su idoneidad depende del tipo de buque y de las características de su comportamiento operativo. Aun así, el modelo **LSTM con filtrado condicionado** destaca como la solución más consistente, al ofrecer el mejor equilibrio entre realismo geográfico y coherencia estadística en la mayoría de los casos analizados.

Tipo de buque	Modelo óptimo
0-Carguero	Se decide el modelo LSTM con filtrado de control ya que presenta: - Mayor parecido geográfico con los datos reales. - Buena coherencia en la distribución de las características evaluadas a pesar de sobreestimar las distancias totales de las trayectorias. Puesto que los cargueros acostumbran a realizar trayectorias extensas, se considera este modelo óptimo ya que el resto de los modelos tienen tendencia a subestimar estas distancias.
1-Petrolero	Se decide el modelo LSTM con filtrado condicionado ya que presenta: - Mayor parecido geográfico con los datos reales. - Buena coherencia en la distribución de las características evaluadas. No se descarta el modelo LSTM con filtrado de control, pero falla en la distribución de distancias generando más trayectorias largas de las encontradas en los datos reales.
2-Pesquero	Se decide el modelo LSTM con filtrado condicionado ya que presenta: - Mayor parecido geográfico con los datos reales. - Buena coherencia en la distribución de las características evaluadas representando correctamente el carácter corto de este tipo de trayectorias. No se descarta el modelo SeqVAE, ya que las trayectorias muestran correctamente el carácter corto y local de los pesqueros, pero falla en la generación de trayectorias de longitud media y larga.
3-Pasajero	Se decide el modelo LSTM con filtrado condicionado ya que presenta: - Mayor parecido geográfico con los datos reales. - Buena coherencia en la distribución de las características evaluada. - Mayor aproximación estadística a los datos reales en cuanto a características generales.
4-Remolcador	Se decide el modelo LSTM con filtrado condicionado ya que presenta: - Mayor parecido geográfico con los datos reales. - Buena coherencia en la distribución de las características evaluadas representando correctamente el carácter corto de este tipo de trayectorias. No se descarta el modelo SeqVAE, ya que las trayectorias muestran correctamente el carácter corto y local de los pesqueros.

Tabla 4-8 Modelo de generación óptimo según el tipo de buque

5 CONCLUSIONES Y LÍNEAS FUTURAS

5.1 Conclusiones

A lo largo del presente Trabajo Fin de Grado se ha abordado el problema de la generación de datos AIS dinámicos sintéticos mediante técnicas de aprendizaje automático, enmarcado dentro de la línea de investigación del CUD-ENM orientada al análisis y modelado del tráfico marítimo.

En primer lugar, se ha llevado a cabo un proceso exhaustivo de preprocesado y análisis de datos AIS reales, permitiendo comprender las características fundamentales de las trayectorias marítimas y establecer una base sólida para el entrenamiento de los modelos. Este proceso ha resultado clave para garantizar la calidad y coherencia de los datos utilizados.

Posteriormente, se han implementado y evaluado dos enfoques principales para la generación de trayectorias sintéticas: un modelo basado en redes neuronales recurrentes (LSTM) y un modelo generativo basado en Autocodificadores Variacionales (SeqVAE). Ambos modelos han demostrado ser capaces de capturar patrones relevantes presentes en los datos reales, aunque con diferencias significativas en su comportamiento y resultados.

El modelo LSTM ha mostrado una mayor capacidad para reproducir la estructura secuencial de las trayectorias y generar recorridos coherentes desde el punto de vista espacial. Tras la aplicación de los filtros, se ha conseguido que las características que definen las trayectorias se asemejen más a los datos reales. No obstante, estas mejoras introducen una reducción en la variabilidad y una mayor concentración espacial de las trayectorias.

Por su parte, el modelo SeqVAE ha permitido generar trayectorias con coherencia estadística en las variables dinámicas. Sin embargo, su capacidad para reproducir con precisión la estructura completa de las trayectorias y la variabilidad observada en los datos reales sigue siendo inferior a la del modelo LSTM en términos de distancias totales de las trayectorias.

La evaluación realizada, combinando análisis estadísticos e inspección visual, ha permitido concluir que los modelos son capaces de reproducir el comportamiento global del tráfico marítimo, diferenciando entre tipos de buque y patrones de navegación. Sin embargo, aún existen desafíos en la generación de trayectorias que reflejen con precisión la complejidad y variabilidad de los datos reales.

En definitiva, este trabajo demuestra el potencial de las técnicas de aprendizaje profundo para la generación de datos AIS dinámicos sintéticos, constituyendo una herramienta de gran utilidad para la ampliación de conjuntos de datos y el desarrollo de futuros sistemas de análisis y predicción en el ámbito marítimo.

5.2 Líneas futuras

En base a las conclusiones del apartado anterior, se proponen a continuación varias líneas futuras para ampliar y continuar los avances realizados en este trabajo:

- Explorar otras técnicas y arquitecturas generativas para la generación de trayectorias sintéticas, con el objetivo de mejorar la diversidad y el realismo de los datos generados. Explorar el uso de ST-Traj GAN comentado en la sección 2.5 como método de generación.
- Mejora del modelado de la variabilidad y del comportamiento dinámico. Sería de interés desarrollar estrategias que permitan reproducir de forma más fiel la diversidad de trayectorias reales, incorporando cambios de rumbo más realistas y evitando la homogeneización de los recorridos generados.
- Desarrollo de métricas de evaluación más completas. Se propone avanzar en la definición de indicadores que permitan evaluar de forma más precisa la similitud entre trayectorias reales y sintéticas, especialmente en términos de estructura espacial y comportamiento dinámico.
- Optimización del coste computacional de los modelos a pesar de que un tiempo elevado de generación no supone una grave limitación, pues la generación de trayectorias sintéticas no está planteada para realizarse en tiempo real.
- Control de la distribución espacial de las trayectorias generadas. Se plantea incorporar mecanismos que permitan forzar o condicionar la distribución geográfica durante el proceso de generación, con el objetivo de evitar concentraciones excesivas y mejorar la representatividad espacial, aunque esto pueda suponer un incremento significativo en el tiempo de generación.
- Mejora del modelo SeqVAE para la generación de trayectorias más largas. Sería de interés adaptar la arquitectura del modelo o su proceso de entrenamiento para permitir la generación de trayectorias de mayor longitud, manteniendo la coherencia temporal y la estabilidad de las secuencias generadas.

6 BIBLIOGRAFÍA

- [1] M. Hidalgo Sánchez, «Generación de datos sintéticos para entrenamiento de sistemas basados en aprendizaje automático supervisado,» TFG Centro Universitario de la Defensa - Escuela Naval Militar, 2025. [En línea]. Available: <https://calderon.cud.uvigo.es/handle/123456789/959>. [Último acceso: Enero 2026].
- [2] J. Fonseca y F. Bacao, «*Tabular and latent space synthetic data generation: a literature review,*» *Journal of Big Data*, vol. 10, nº 115, 2023.
- [3] M. Goyal y Q. H. Mahmoud, «*A systematic review of synthetic data generation techniques using generative AI,*» *Electronics*, vol. 13, nº 17, p. 3509, 2024.
- [4] Y. Lu, M. Shen, H. Wang, X. Wang, C. Van Rechem, T. Fu y W. Wei, «*Machine Learning for Synthetic Data Generation: A Review,*» arxiv preprint, 4 Abril 2025.
- [5] M. Rodelgo Lacruz, B. Barragáns Martínez, N. Fernández García, G. Higgins, P. Sendín Raña y A. Suárez García, «*Real-time AIS data analysis for anomaly detection in the maritime environment,*» de *17th NATO Operations Research and Analysis (OR&A) Conference*, Laurel, Maryland (USA), 2023.
- [6] D. P. Kingma y M. Welling, «*An Introduction to Variational Autoencoders,*» *Foundations and Trends in Machine Learning*, vol. 12, nº 4, pp. 307-392, 2019.
- [7] D. P. Kingma y M. Welling, «*Auto-Encoding Variational Bayes,*» 10 Diciembre 2022. [En línea]. Available: <https://doi.org/10.48550/arXiv.1312.6114>.
- [8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville y Y. Bengio, «*Generative Adversarial Networks,*» de *Advances in Neural Information Processing Systems (Vol 27)*, Montreal, Canadá, 2014.
- [9] A. Figueira y B. Vaz, «*Survey on Synthetic Data Generation, Evaluation Methods and GANs,*» 2022. [En línea]. Available: <https://doi.org/10.3390/math10152733>. [Último acceso: Enero 2026].
- [10] J. Gui, Z. Sun, Y. Wen, D. Tao y J. Ye, «*A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications,*» arXiv.org, 2020.

- [11] S. Hochreiter y J. Schmidhuber, «*Long Short-Term Memory*,» *Neural Computation*, vol. 9, p. 1735–1780, <https://doi.org/10.1162/neco.1997.9.8.1735>, 1997.
- [12] F. Althé y A. de la Fortelle, «*An LSTM Network for Highway Trajectory Prediction*,» de 2017 *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC) Workshops*, <https://doi.org/10.1109/ITSC.2017.8317913>, 16-19 Octubre 2017.
- [13] K. Kaur y Manisha, «*A review on large language models: Architectures, applications, challenges and future directions*,» *International Journal for Multidisciplinary Research (IJFMR)*, vol. 7, nº 6, Diciembre 2025.
- [14] S. Han, M. Wang, J. Zhang, D. Li y J. Duan, «*A review of large language models*,» *MDPI, Electronics*, vol. 12, nº 24, 21 Diciembre 2024.
- [15] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle y A. P. Sales, «*Generation and evaluation of synthetic patient data*,» *BMC Medical Research Methodology*, vol. 20, nº 108, 7 Mayo 2020.
- [16] F. J. Anscombe, «*Graphs in Statistical Analysis*,» *The American Statistician*, vol. 27, nº 1, pp. 17-21, Febrero 1973.
- [17] *International Association of Marine Aids to Navigation and Lighthouse Authorities (IALA)*, «*Guía 1082: Perspectiva general del Sistema de Identificación Automática (AIS)*,» IALA, Saint-Germain-en-Laye, Francia, 2011.
- [18] K. Jaskólski, Ł. Marchel, A. Felski, M. Jaskólski y M. Specht, «*Automatic Identification System (AIS) Dynamic Data Integrity Monitoring and Trajectory Tracking Based on the Simultaneous Localization and Mapping (SLAM) Process Model*,» *Sensors*, vol. 21, nº 8430, p. <https://doi.org/10.3390/s21248430>, 2021.
- [19] F. Heymann, T. Noack y P. Banys, «*Plausibility analysis of navigation related AIS parameter based on time series*,» de *European Navigation Conference*, Neustrelitz, Germany, 2013.
- [20] G. Rodríguez Casajús, «*Predicción de tipo de buque utilizando datos AIS y técnicas de inteligencia artificial*,» TFG Centro Universitario de la Defensa - Escuela Naval Militar, 2022. [En línea]. Available: <https://calderon.cud.uvigo.es/handle/123456789/646>. [Último acceso: 09 Febrero 2026].
- [21] J. Roldán Calvo, «*Predicción de tipo de buque utilizando representaciones de trayectorias y técnicas de inteligencia artificial*,» TFG Centro Universitario de la Defensa - Escuela Naval Militar, 2024. [En línea]. Available: <https://calderon.cud.uvigo.es/handle/123456789/813>. [Último acceso: 09 Febrero 2026].
- [22] F. J. Pavón Jarava, «*Mejora de la eficiencia computacional de un sistema de predicción de tipo de buque utilizando parámetros de trayectorias y técnicas de IA*,» TFG Centro Univeritario de la Defensa - Escuela Naval Militar, 2025. [En línea]. Available: <https://calderon.cud.uvigo.es/handle/123456789/951>. [Último acceso: 09 Febrero 2026].
- [23] M. Xuebin, D. Zinan y Z. Xiaoyan, «*ST-TrajGAN: A synthetic trajectory generation algorithm for privacy preservation*,» *Future Generation Computer Systems*, vol. 161, pp. 226-238, December 2024.
- [24] Project Jupyter, «*The JupyterLab Interface*,» Jupyter Labs, [En línea]. Available: <https://jupyterlab.readthedocs.io/en/stable/user/interface.html>.
- [25] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing y Jupyter

- Development Team, « *Jupyter Notebooks – a publishing format for reproducible computational workflows*,» Positioning and Power in Academic Publishing. IOS Press., 2016. [En línea]. Available: <https://eprints.soton.ac.uk/403913/1/STAL9781614996491-0087.pdf>. [Último acceso: Febrero 13 2026].
- [26] J. F. Pimentel, L. Murta, V. Braganholo y J. Freire, «*A large-scale study about quality and reproducibility of Jupyter Notebooks*,» *Proceedings of MSR*, 2019.
- [27] A. Rule, A. Birmingham, C. Zuniga, I. Altintas, S.-C. Huang, K. Rob, N. Moshiri, M. H. Nguyen, S. B. Rosenthal, F. Pérez y P. W. Rose, «*Ten simple rules for writing and sharing computational analyses in Jupiter Notebooks*,» *PLOS Computational Biology*, vol. 14, n° 7, 25 Julio 2019.
- [28] Pandas Development Team, «*Pandas Documentation*,» 2024. [En línea]. Available: <https://pandas.pydata.org/docs/>. [Último acceso: 16 Febrero 2026].
- [29] The NumPy Developers , «*NumPy Documentation*,» 2024. [En línea]. Available: <https://numpy.org/doc/>. [Último acceso: 16 Febrero 2026].
- [30] The Matplotlib Development Team, «*Matplotlib Documentation*,» 2024. [En línea]. Available: <https://matplotlib.org/stable/>. [Último acceso: 16 Febrero 2026].
- [31] The Folium Developers , «*Folium Documentation*,» 2024. [En línea]. Available: <https://python-visualization.github.io/folium/latest/>. [Último acceso: 02 Marzo 2026].
- [32] The PyTorch Developers, «*PyTorch Documentation*,» 2019. [En línea]. Available: <https://docs.pytorch.org/docs/stable/>. [Último acceso: 02 Marzo 2026].
- [33] The TensorFlow Developers, «*TensorFlow documentation*,» 2016. [En línea]. Available: <https://www.tensorflow.org/?hl=es-419>. [Último acceso: 02 Marzo 2026].
- [34] The Apache Arrows Developers, «*The Apache Arrows Documentation*,» 2016. [En línea]. Available: <https://arrow.apache.org/docs/>. [Último acceso: 03 Marzo 2026].
- [35] The GeoPandas Developers, «*GeoPandas Documentation*,» 2024. [En línea]. Available: <https://geopandas.org/en/stable/>. [Último acceso: 03 Marzo 2026].
- [36] *Natural Earth*, «Admin 0 – Countries (1:110m),» [En línea]. Available: https://naturalearth.s3.amazonaws.com/110m_cultural/ne_110m_admin_0_countries.zip. [Último acceso: 24 Marzo 2026].
- [37] Y. Shi, C. Long, X. Yang y M. Deng, «*Abnormal Ship Behavior Detection Based on AIS Data*,» *Applied Sciences*, vol. 12, n° 9, p. <https://doi.org/10.3390/app12094635>, 2022.
- [38] UK Data Service; ADR UK, «*The role of synthetic data in research: Benefits, risks and governance*,» Zenodo, Ginebra, 10 Abril 2025.
- [39] M. Miletic y M. Sariyar, «*Challenges of Using Synthetic Data Generation Methods for Tabular Microdata*,» *Applied Sciences*, vol. 14, n° 14, p. 5975, 9 Julio 2024.
- [40] F. Pérez y B. E. Granger, «*IPython: A System for Interactive Scientific Computing*,» *Computing in Science & Engineering*, vol. 9, n° 3, pp. 21-29, 2007.
- [41] S. Konstantin, S. Cordelia y A. Karteeq, «*How good is my GAN?*,» *European Conference on Computer Vision*, 2018.

ANEXO I: IMPLICACIONES SOCIALES, Y/O ECONÓMICAS, Y/O AMBIENTALES

La generación de datos sintéticos mediante técnicas de inteligencia artificial está teniendo un impacto creciente en múltiples sectores, al permitir paliar la escasez de datos reales, facilitar el desarrollo de investigaciones, preservar la privacidad y reducir los costes asociados a la obtención de información. En el contexto del presente Trabajo Fin de Grado, centrado en la generación de trayectorias AIS dinámicas sintéticas, estas ventajas adquieren especial relevancia en el ámbito marítimo.

Desde una perspectiva **social**, los datos sintéticos permiten ampliar el acceso a conjuntos de datos que, en muchos casos, están limitados por cuestiones de confidencialidad o seguridad. En el ámbito marítimo, los datos AIS pueden contener información sensible relacionada con rutas, patrones de navegación o actividades de determinados buques. La generación de datos sintéticos posibilita trabajar con información representativa sin necesidad de exponer directamente estos datos reales, favoreciendo así el desarrollo de investigaciones y modelos de aprendizaje automático sin comprometer la seguridad de la información.

En relación con los aspectos de **seguridad**, el uso de datos sintéticos permite desarrollar y entrenar sistemas de análisis del tráfico marítimo en entornos controlados. Esto resulta especialmente relevante en aplicaciones como la detección de anomalías, la predicción de trayectorias o la monitorización del tráfico, donde el acceso a datos reales puede estar restringido. La disponibilidad de datos sintéticos contribuye a mejorar la robustez de estos sistemas, permitiendo su validación sin depender exclusivamente de datos operativos reales.

Desde el punto de vista **económico**, la generación de datos sintéticos supone un ahorro significativo de recursos, ya que reduce la necesidad de realizar esfuerzos de recopilación, almacenamiento y tratamiento de datos reales, procesos que suelen implicar elevados costes tanto en tiempo como en medios técnicos. En este trabajo, las trayectorias AIS sintéticas permiten ampliar el volumen de datos disponible para el entrenamiento de modelos, mejorando su capacidad de generalización y optimizando los recursos empleados en el desarrollo de soluciones basadas en aprendizaje automático.

En cuanto a los aspectos **ambientales**, es importante considerar que el entrenamiento de modelos de inteligencia artificial conlleva un consumo energético significativo. Por ello, resulta necesario avanzar hacia modelos más eficientes y hacia un uso responsable de los recursos computacionales, buscando un equilibrio entre el desarrollo tecnológico y la sostenibilidad.

Por último, desde el punto de vista **industrial**, el desarrollo de este trabajo implica la aplicación de técnicas avanzadas de ingeniería para el diseño, implementación y validación de modelos generativos basados en aprendizaje profundo. Este tipo de soluciones tiene un alto potencial de aplicación en sistemas reales de análisis del tráfico marítimo, contribuyendo a la mejora de procesos de monitorización, planificación y toma de decisiones en el sector.

De este modo, el presente trabajo pone de manifiesto que la generación de datos sintéticos no solo constituye una herramienta técnica de gran utilidad, sino que también presenta implicaciones relevantes en ámbitos sociales, económicos, ambientales y de seguridad, que deben ser tenidas en cuenta en su desarrollo y aplicación.

ANEXO II: REFLEXIONES ÉTICAS Y SOCIALES

El desarrollo de técnicas de generación de datos sintéticos mediante inteligencia artificial plantea una serie de cuestiones éticas y sociales que deben ser consideradas en el contexto de su aplicación, especialmente en ámbitos sensibles como el marítimo.

En primer lugar, uno de los principales aspectos éticos asociados a este tipo de tecnologías es el equilibrio entre la **utilidad de los datos y la protección de la información sensible**. En el ámbito del presente trabajo, la generación de trayectorias AIS sintéticas permite trabajar con datos representativos del tráfico marítimo sin necesidad de utilizar directamente datos reales, lo que contribuye a preservar la confidencialidad de la información. Sin embargo, es importante tener en cuenta que, si los modelos no están adecuadamente diseñados, podrían llegar a reproducir patrones específicos de los datos originales, comprometiendo indirectamente dicha privacidad. Por ello, resulta fundamental evaluar no solo la calidad de los datos generados, sino también su grado de desvinculación respecto a los datos reales.

Otro aspecto relevante es la **fiabilidad de los datos sintéticos** y su uso en sistemas críticos. En aplicaciones como la vigilancia marítima, la detección de anomalías o la predicción de trayectorias, la utilización de datos no representativos o con sesgos puede derivar en modelos poco robustos o incluso en decisiones incorrectas. En este sentido, el presente trabajo pone de manifiesto que, aunque los modelos generativos son capaces de reproducir el comportamiento global de las trayectorias, aún presentan limitaciones en la representación de la variabilidad real, lo que refuerza la necesidad de utilizar estos datos de forma complementaria y no como sustituto absoluto de los datos reales.

Desde una perspectiva social, la generación de datos sintéticos contribuye a la **democratización del acceso a la información**, permitiendo que investigadores y desarrolladores puedan trabajar con datos similares a los reales sin restricciones legales o de seguridad. Esto favorece el avance científico y tecnológico, reduciendo barreras de acceso y fomentando la innovación. No obstante, también plantea el riesgo de un uso indebido de estas tecnologías, por ejemplo, en la generación de información artificial con fines maliciosos o en la simulación de comportamientos no reales con intenciones engañosas.

Asimismo, debe considerarse el impacto relacionado con la **responsabilidad en el desarrollo y uso de sistemas basados en inteligencia artificial**. El diseño de modelos generativos implica la toma de decisiones sobre el tratamiento de los datos, la selección de variables y la evaluación de resultados, lo que introduce posibles fuentes de sesgo. Es responsabilidad del desarrollador garantizar que estos sistemas sean lo más transparentes, robustos y fiables posible, así como reconocer sus limitaciones.

Por otro lado, el uso de estas tecnologías también plantea cuestiones relacionadas con la **sostenibilidad digital**, ya que el entrenamiento de modelos de aprendizaje profundo requiere un consumo considerable de recursos computacionales. Este aspecto obliga a reflexionar sobre la necesidad de desarrollar soluciones más eficientes y de optimizar el uso de los recursos disponibles, buscando un equilibrio entre el avance tecnológico y su impacto energético.

En resumen, la generación de datos sintéticos representa una herramienta de gran potencial, pero cuyo uso debe ir acompañado de una reflexión ética que tenga en cuenta la privacidad, la fiabilidad, la seguridad y el impacto social de estas tecnologías. En este trabajo se ha adoptado una aproximación responsable, reconociendo tanto las ventajas como las limitaciones de los modelos desarrollados, y destacando la importancia de un uso adecuado de los datos sintéticos en aplicaciones reales.

ANEXO III: RESULTADOS PRIMERA PRUEBA DE LSTM PARA PETROLEROS, PASAJEROS Y REMOLCADORES

El presente anexo recoge los resultados correspondientes a la primera versión del modelo LSTM aplicada a distintos tipos de buque, concretamente petroleros, pasajeros y remolcadores. Estos resultados se incluyen como material complementario con el objetivo de mostrar el comportamiento inicial del modelo antes de la aplicación de las mejoras y filtrados descritos en el capítulo 4.

El análisis de estas trayectorias permite identificar las principales limitaciones del modelo en su fase inicial, especialmente en lo relativo a la coherencia espacial, la escala de los recorridos y la representación de los patrones de navegación característicos de cada tipo de buque. De este modo, este anexo proporciona un contexto adicional que facilita la comprensión de las decisiones adoptadas durante el proceso de ajuste del modelo y la evolución de los resultados obtenidos.

Petroleros

A continuación, se analiza el comportamiento del modelo para el caso de los buques petroleros, caracterizados, al igual que los cargueros, por trayectorias de media y larga distancia asociadas a rutas comerciales entre terminales energéticas.

Tal y como se ha observado en el apartado 4.1, las trayectorias generadas mediante el modelo LSTM presentan una distribución geográfica amplia y coherente, manteniendo la naturaleza global de este tipo de tráfico, aunque con una cierta concentración en determinadas regiones.

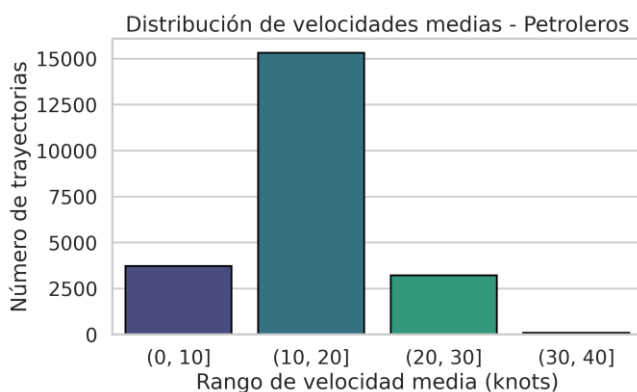


Figura A3-1 Distribución velocidades reales Petroleros

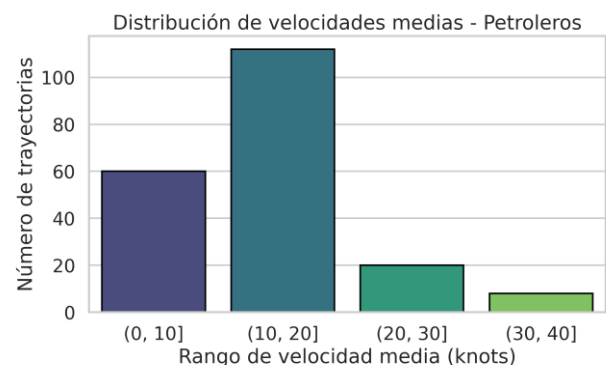


Figura A3-2 Distribución velocidades LSTM Petroleros

Desde el punto de vista estadístico, y como se aprecia en las **gráficas de distribución de velocidades, duración y distancia** (Figura A3-1, Figura A3-2, Figura A3-3, Figura A3-4, Figura A3-5 y Figura A3-6), el modelo reproduce adecuadamente los rangos predominantes de velocidad, con una clara concentración en el intervalo de 10 a 20 nudos, en línea con los datos reales. Sin embargo, al igual que en el caso de los cargueros, se observa una menor presencia en los rangos de velocidad más elevados, lo que indica una tendencia del modelo a concentrarse en los valores más frecuentes.

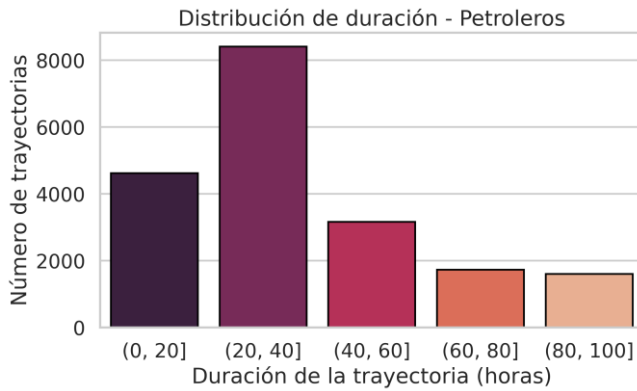


Figura A3-3 Distribución tiempo reales Petroleros

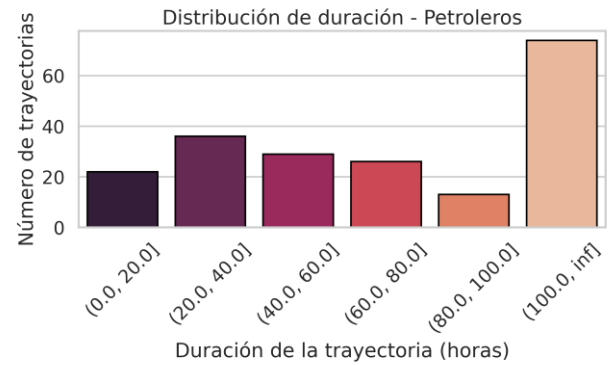


Figura A3-4 Distribución tiempo LSTM Petroleros

En cuanto a la **duración de las trayectorias**, los datos reales presentan una distribución relativamente amplia, con predominio de trayectorias de duración media. No obstante, en las trayectorias generadas se aprecia una tendencia a producir trayectorias de mayor duración, lo que se traduce en valores medios superiores.

De forma coherente, en la **distribución de distancias** (Figura A3-5 y Figura A3-6) puede apreciarse que las trayectorias generadas presentan una mayor concentración en los intervalos de distancia más elevados, en comparación con los datos reales. Este comportamiento está directamente relacionado con la mayor duración de las trayectorias sintéticas, lo que conlleva una mayor distancia acumulada.

Por último, se analiza la **distribución geográfica de las trayectorias** (Figura A3-7 y Figura A3-8), clasificada según el cuadrante predominante del desplazamiento. En los datos reales se detecta una distribución relativamente equilibrada entre los distintos cuadrantes, lo que refleja la diversidad de rutas de los petroleros. En las trayectorias generadas mediante el modelo LSTM también se mantienen trayectorias en los distintos cuadrantes, aunque con ciertas variaciones en las proporciones, en línea con lo observado en el análisis visual.

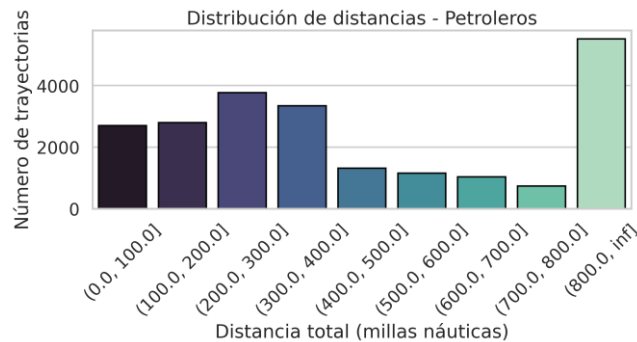


Figura A3-5 Distribución distancias reales Petroleros

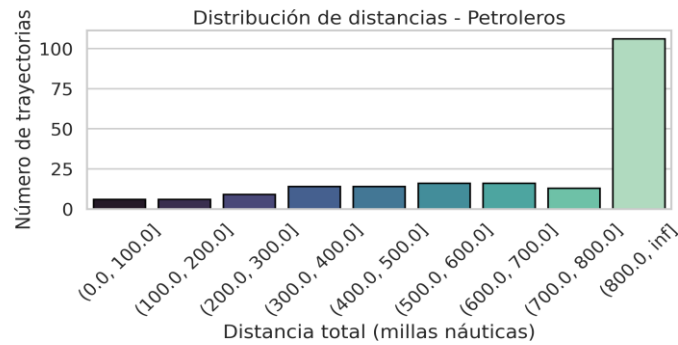


Figura A3-6 Distribución distancias LSTM Petroleros

Estas diferencias pueden explicarse, en parte, por las características geográficas de determinadas rutas marítimas, especialmente aquellas próximas a zonas costeras o estrechos, donde las rutas transitables son más estrechas. Durante el proceso de generación se aplica un filtrado que elimina aquellas trayectorias que presentan puntos situados sobre tierra firme, lo que puede provocar la eliminación de algunas trayectorias y afectar ligeramente a la distribución final en determinados cuadrantes.

En términos globales, los resultados muestran que el modelo LSTM reproduce adecuadamente el tipo de trayectoria característico de los petroleros, manteniendo su naturaleza de recorridos de media y larga distancia. No obstante, al igual que en el caso de los cargueros, se observa una tendencia a sobreestimar la duración y la distancia de las trayectorias, lo que indica una ligera desviación en la escala de los recorridos generados.

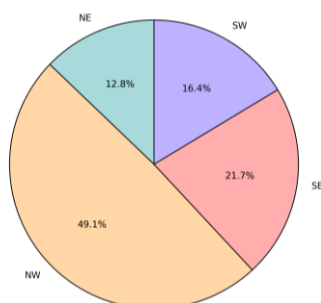


Figura A3-7 Distribución geográfica reales Petroleros

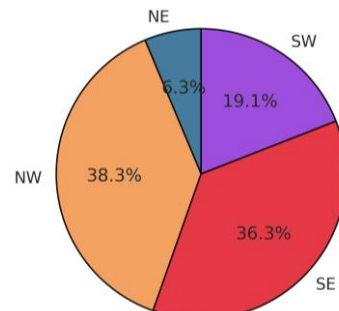


Figura A3-8 Distribución geográfica LSTM Petroleros

Pasajeros

A continuación, se analiza el comportamiento del modelo para el caso de los buques de pasajeros, cuyas trayectorias se caracterizan por recorridos cortos o de media distancia, generalmente asociados a rutas regionales y navegación en entornos costeros.

Tal y como se ha observado en el apartado 4.1, las trayectorias generadas mediante el modelo LSTM mantienen una estructura coherente, aunque con una extensión algo superior a la esperada para este tipo de tráfico.

las **gráficas de distribución de velocidades** (Figura A3-9 y Figura A3-10) se aprecia que el modelo reproduce adecuadamente la concentración en el intervalo de 0 a 20 nudos, con especial presencia en el rango inferior, lo que refleja correctamente la combinación de navegación y maniobras característica de estos buques.

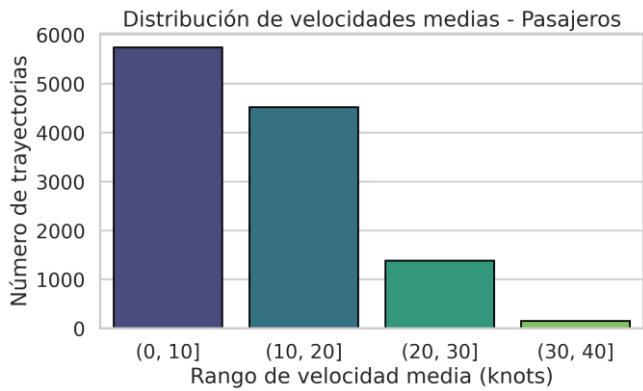


Figura A3-9 Distribución velocidades reales Pasajeros

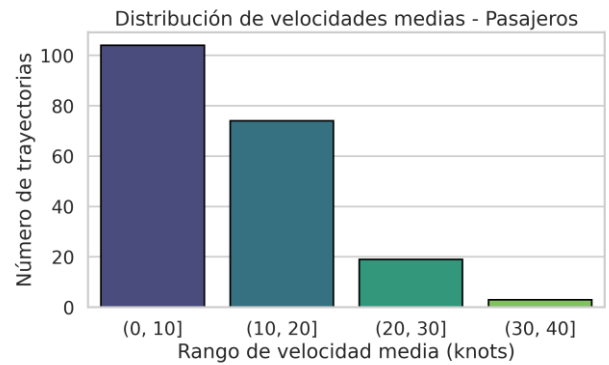


Figura A3-10 Distribución velocidades LSTM Pasajeros

Sin embargo, en las **distribuciones de duración y distancia** (Figura A3-11, Figura A3-12, Figura A3-13 y Figura A3-14) se observa una mayor dispersión hacia intervalos superiores respecto a los datos reales. En particular, las trayectorias generadas presentan una mayor presencia de recorridos de duración media y larga, lo que se traduce en distancias acumuladas más elevadas.

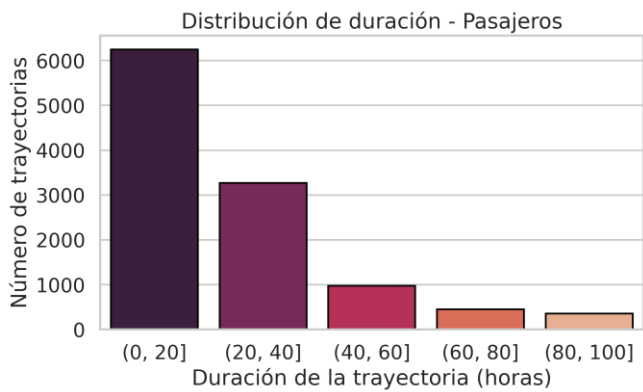


Figura A3-11 istribución tiempo reales Pasajeros

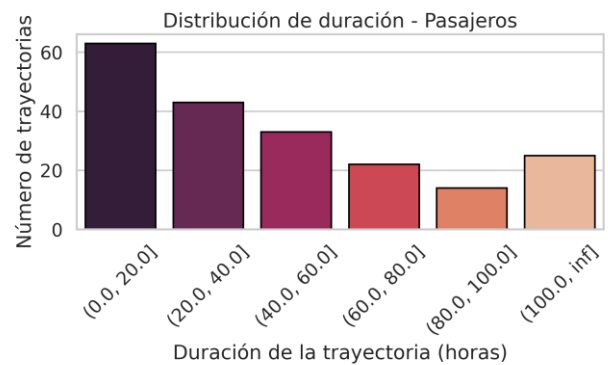


Figura A3-12 istribución tiempo LSTM Pasajeros

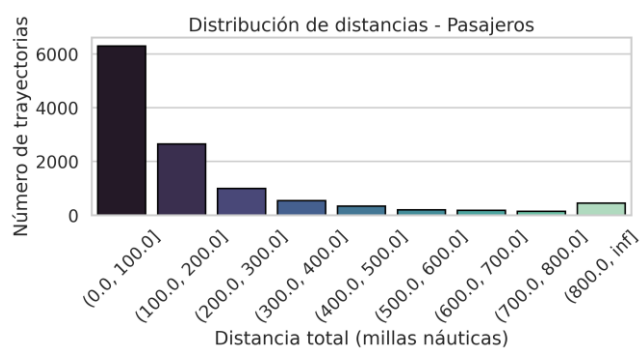


Figura A3-13 Distribución distancia reales Pasajeros

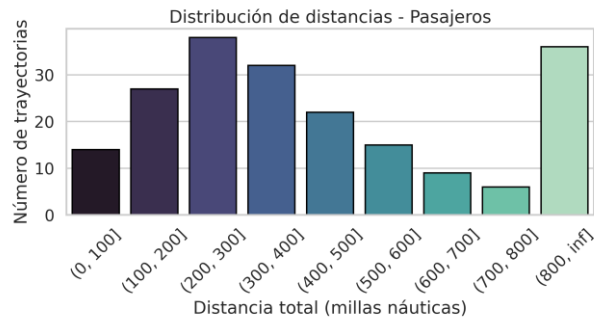


Figura A3-14 Distribución distancia LSTM Pasajeros

En cuanto a la **distribución geográfica** (Figura A3-15 y Figura A3-16), las trayectorias generadas mantienen presencia en los distintos cuadrantes, aunque con una mayor concentración en los cuadrantes noroeste (NW) y sureste (SE). Estas variaciones, al igual que en los casos anteriores, están condicionadas en parte por el proceso de filtrado aplicado durante la validación, especialmente relevante en trayectorias próximas a la costa.

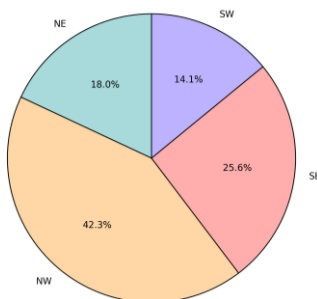


Figura A3-15 Distribución geográfica reales Pasajeros

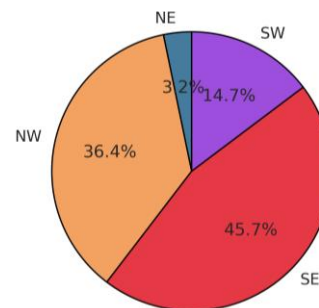


Figura A3-16 Distribución geográfica LSTM Pasajeros

En conjunto, el modelo LSTM reproduce de forma adecuada la naturaleza regional de las trayectorias de pasajeros, capturando sus rangos de velocidad característicos, aunque con una tendencia a generar trayectorias más largas y con mayor dispersión que las observadas en los datos reales.

Remolcadores

A continuación, se analiza el comportamiento del modelo para el caso de los remolcadores, cuyas trayectorias se caracterizan por ser muy cortas, locales y asociadas a entornos portuarios o zonas próximas a la costa.

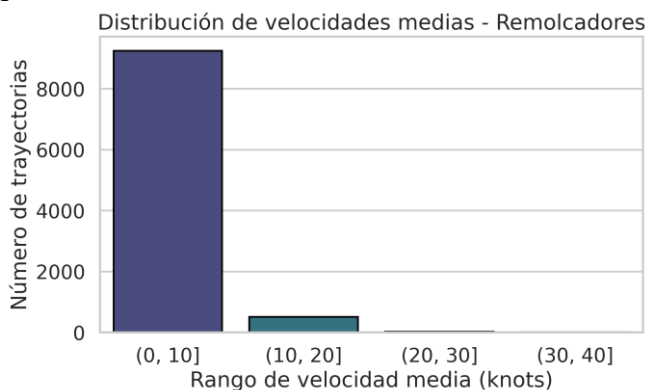


Figura A3-17 Distribución velocidades reales Remolcadores

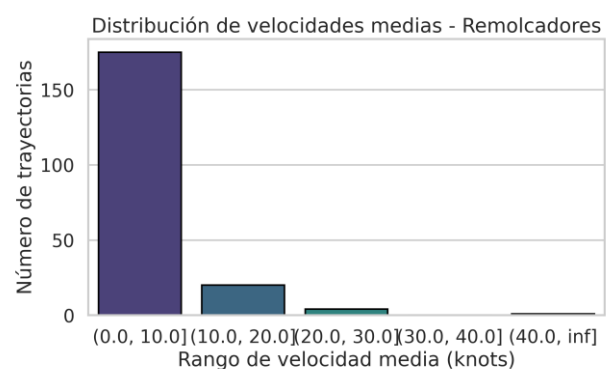


Figura A3-18 Distribución velocidades LSTM Remolcadores

Tal y como se ha observado en el apartado 4.1, las trayectorias generadas mediante el modelo LSTM mantienen en gran medida este carácter local, aunque con una ligera tendencia a ampliar la extensión de los recorridos.

En las **gráficas de distribución de velocidades** (Figura A3-17 y Figura A3-18) se aprecia una clara concentración en el intervalo de 0 a 10 nudos, en línea con los datos reales, lo que indica que el modelo captura correctamente el patrón de navegación a baja velocidad propio de este tipo de buques.

Sin embargo, al analizar las **distribuciones de duración y distancia** (Figura A3-19, Figura A3-20, Figura A3-21 y Figura A3-22), se observa una mayor dispersión hacia valores superiores en comparación con los datos reales. Esto se traduce en trayectorias sintéticas que, aunque siguen siendo relativamente cortas, presentan una duración y una distancia acumulada algo mayores de lo esperado.

Como en los casos anteriores, este comportamiento puede estar influido por el procedimiento probabilístico utilizado para la reconstrucción de la componente temporal, que introduce pequeñas variaciones acumulativas en los intervalos entre puntos.

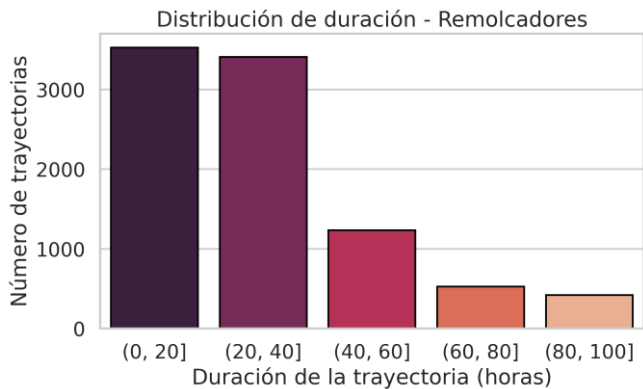


Figura A3-19 Distribución tiempo reales Remolcadores

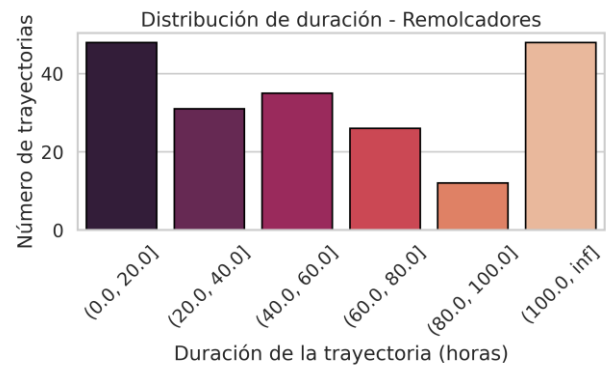


Figura A3-20 Distribución tiempo LSTM Remolcadores

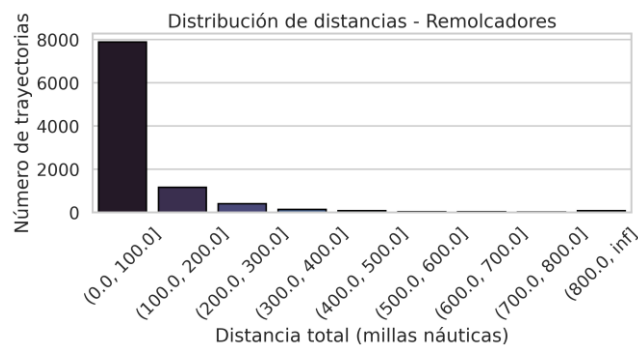


Figura A3-21 Distribución distancia reales Remolcadores

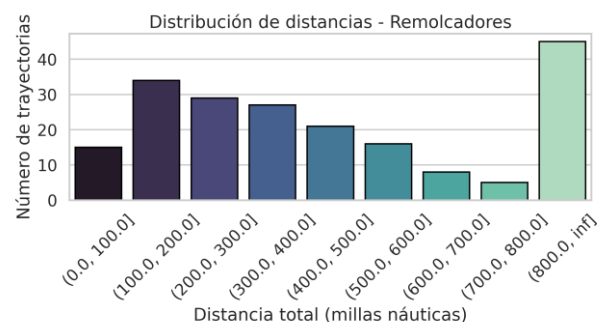


Figura A3-22 Distribución distancias LSTM Remolcadores

En cuanto a la **distribución geográfica** (Figura A3-23 y Figura A3-24), las trayectorias generadas mantienen presencia en los distintos cuadrantes, reflejando la diversidad de movimientos locales, aunque con ligeras variaciones en las proporciones respecto a los datos reales.

Considerando todos los resultados, el modelo LSTM reproduce adecuadamente el carácter local y de baja velocidad de los remolcadores, diferenciándolo de otros tipos de buque. No obstante, persiste una tendencia a generar trayectorias ligeramente más largas, tanto en duración como en distancia, lo que introduce una desviación moderada respecto al comportamiento real.

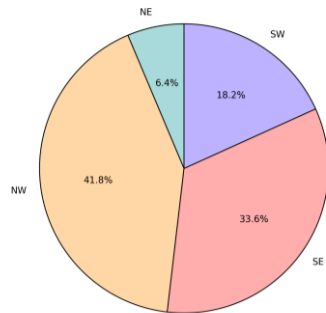


Figura A3-23 Distribución geográfica reales Remolcadores

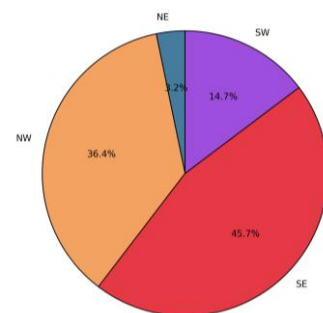


Figura A3-24 Distribución geográfica LSTM Remolcadores

ANEXO IV: RESULTADOS DEL FILTRADO DE CONTROL PARA LSTM

Cargueros

A continuación, se analiza el comportamiento del modelo LSTM tras la aplicación del filtrado de control para el caso de los buques cargueros, a partir de las distribuciones obtenidas.

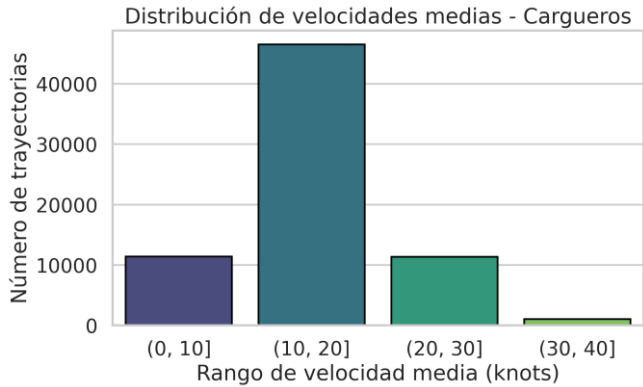


Figura A4-1 Distribución velocidades reales Cargueros

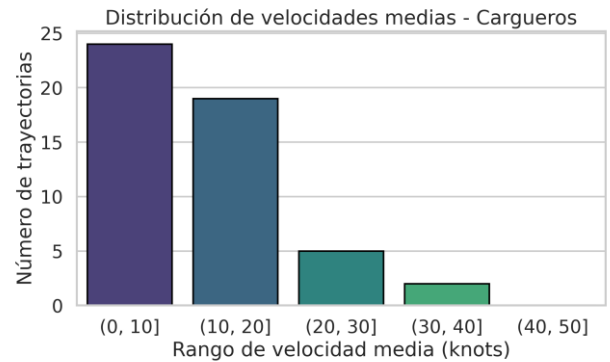


Figura A4-2 Distribución velocidades LSTM con filtrado de control para Cargueros

En las **gráficas de distribución de velocidades medias** (Figura A4-1 y Figura A4-2) se observa que las trayectorias generadas mantienen la forma general de los datos reales, con una clara concentración en el intervalo de 10 a 20 nudos. Aunque existe una ligera redistribución entre intervalos, el modelo reproduce adecuadamente el rango predominante de velocidades, lo que indica que esta variable no se ve afectada de forma significativa por el filtrado.

En cuanto a la **distribución de la duración de las trayectorias** (Figura A4-3 y Figura A4-4), se aprecia una mayor presencia de trayectorias en intervalos de duración elevada en comparación con los datos reales. No obstante, la distribución mantiene una forma coherente, con presencia en todos los intervalos y una estructura similar a la observada en los datos originales.

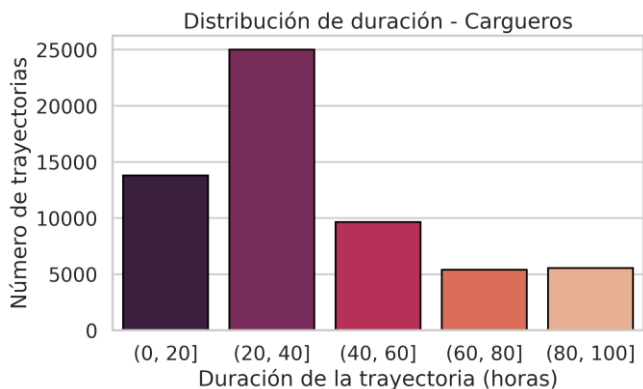


Figura A4-3 Distribución tiempo reales Cargueros

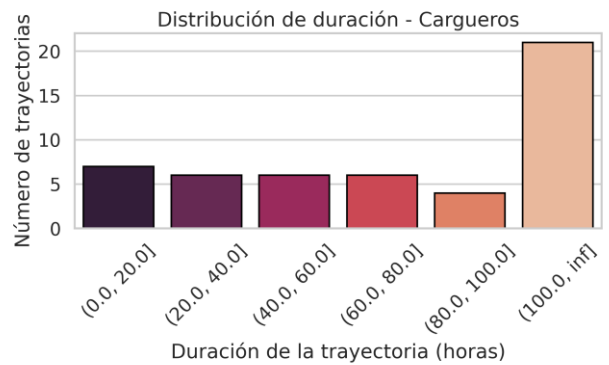


Figura A4-4 Distribución tiempo LSTM con filtrado de control para Cargueros

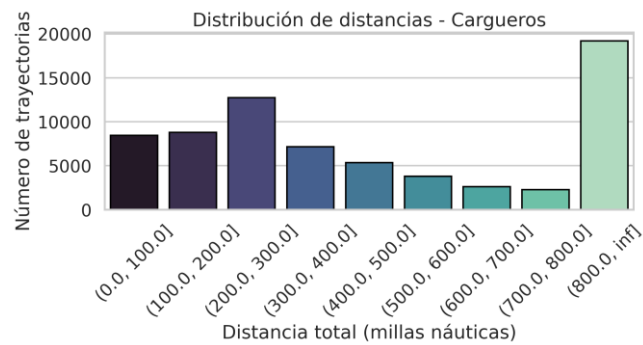


Figura A4-5 Distribución distancias reales Cargueros

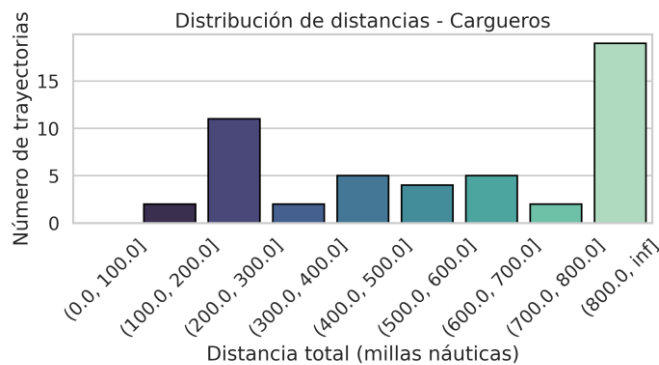


Figura A4-6 Distribución distancias LSTM con filtrado de control para Cargueros

La **distribución de distancias** (Figura A4-5 y Figura A4-6) muestra el efecto más relevante del filtrado. Mientras que en los datos reales existe una distribución más equilibrada, en las trayectorias generadas se observa una concentración en intervalos de distancia media-alta, con una reducción de los valores más extremos. Esto indica que el filtrado ha limitado la generación de trayectorias excesivamente largas, aunque persiste una ligera tendencia hacia distancias superiores a las reales

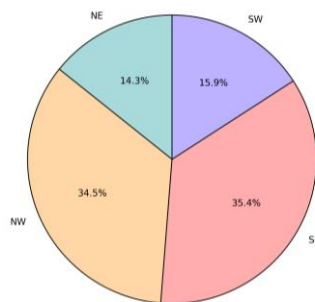


Figura A4-7 Distribución geográfica reales Cargueros

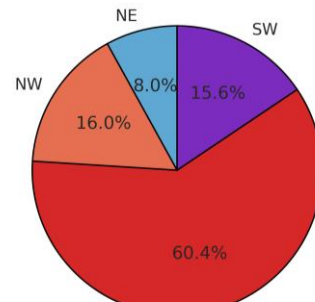


Figura A4-8 Distribución geográfica LSTM con filtrado de control para Cargueros

Por último, en la **distribución geográfica por cuadrantes** (Figura A4-7 y Figura A4-8), se observa que las trayectorias reales presentan una distribución relativamente equilibrada, con cierta predominancia en los cuadrantes NW y SE. En cambio, las trayectorias generadas muestran una mayor concentración en el cuadrante SE, con una menor presencia en otros cuadrantes, especialmente en el NE. Esto refleja una reducción en la dispersión direccional de las trayectorias generadas.

En términos globales, las gráficas muestran que el modelo LSTM, tras el filtrado, mantiene la capacidad de reproducir las características principales de las trayectorias de los cargueros, especialmente en lo relativo a la velocidad. No obstante, persisten ciertas diferencias en la distribución de duración,

distancia y orientación geográfica, lo que indica una ligera desviación en la escala y en la dispersión espacial de los recorridos generados.

Petroleros

Tras la aplicación del filtrado sobre el método de generación del modelo LSTM, se analizan las distribuciones resultantes para el caso de los buques petroleros, comparándolas con los datos reales.

En las gráficas de **distribución de velocidades medias** (Figura A4-9 y Figura A4-10), se observa que el modelo mantiene una buena aproximación a los datos reales, con una clara predominancia del intervalo 10–20 nudos, característico de este tipo de buque. El filtrado no introduce cambios significativos en esta variable, lo que indica que la velocidad no se ve afectada de forma relevante por la eliminación de trayectorias.

En cuanto a la **distribución de la duración** de las trayectorias (Figura A4-11 y Figura A4-12), se aprecia un ajuste más cercano a los datos reales tras el filtrado. Aunque sigue existiendo presencia de trayectorias de mayor duración, la distribución resultante muestra una reducción en los valores extremos, aproximándose mejor al intervalo predominante de duración media observado en los datos reales.

Este efecto se refleja de forma más clara en la **distribución de distancias** (Figura A4-13 y Figura A4-14). Mientras que en los datos reales existe una distribución más equilibrada entre distintos intervalos, en las trayectorias corregidas se reduce notablemente la presencia de recorridos excesivamente largos, eliminando en gran medida los valores más alejados. Como resultado, la distribución de distancias se aproxima mejor a la estructura observada en los datos reales, corrigiendo la tendencia previa del modelo a generar trayectorias sobredimensionadas.

Por último, en la **distribución geográfica por cuadrantes** (Figura A4-15 y Figura A4-16), se observa que las trayectorias generadas mantienen presencia en los distintos cuadrantes, aunque persiste cierta concentración en el cuadrante sureste (SE). No obstante, el filtrado no altera de forma significativa este patrón, lo que indica que su impacto se centra en la escala de las trayectorias más que en su orientación.

En conjunto, los resultados muestran que el filtrado aplicado permite corregir eficazmente la sobreestimación en distancia y duración, mejorando el ajuste del modelo LSTM a los datos reales sin afectar de forma significativa a la distribución de velocidades ni a la estructura general de las trayectorias.

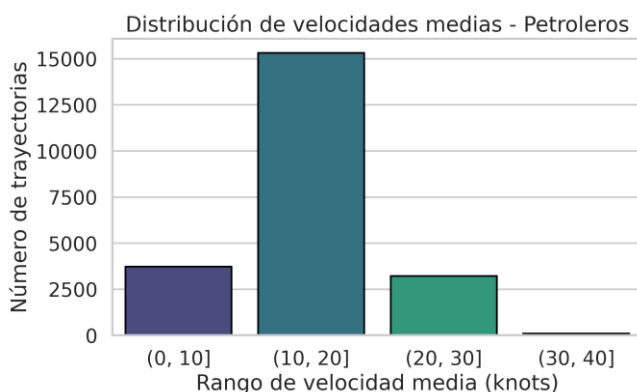


Figura A4-9 Distribución velocidades reales Petroleros

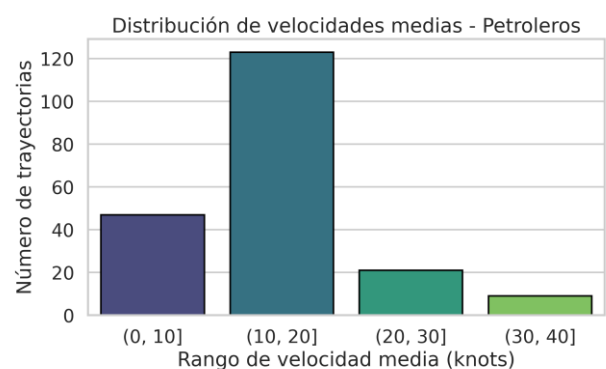


Figura A4-10 Distribución velocidades LSTM con filtrado de control para Petroleros

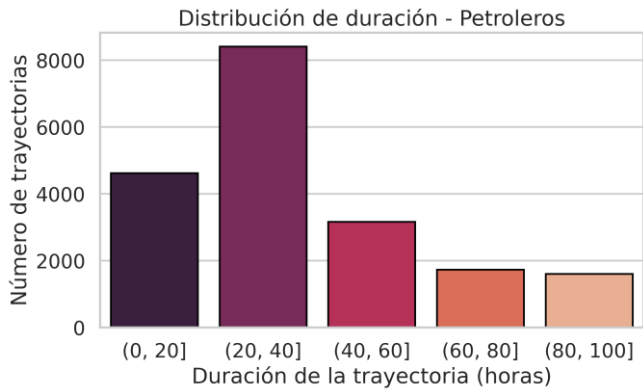


Figura A4-11 Distribución tiempo reales Petroleros

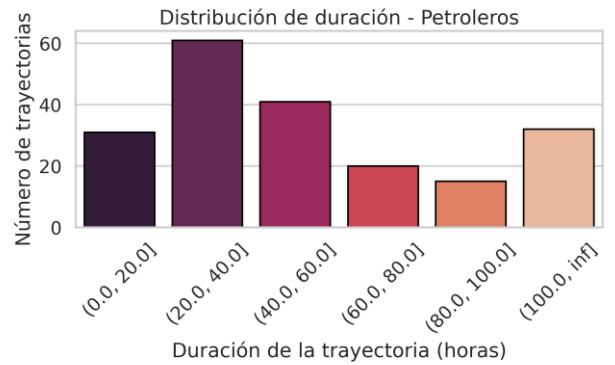


Figura A4-12 Distribución tiempo LSTM con filtrado de control para Petroleros

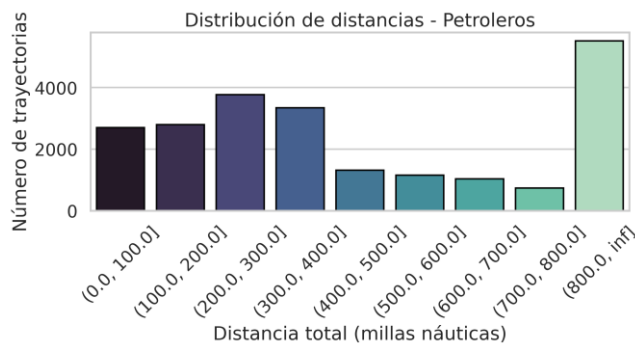


Figura A4-13 Distribución distancias reales Petroleros

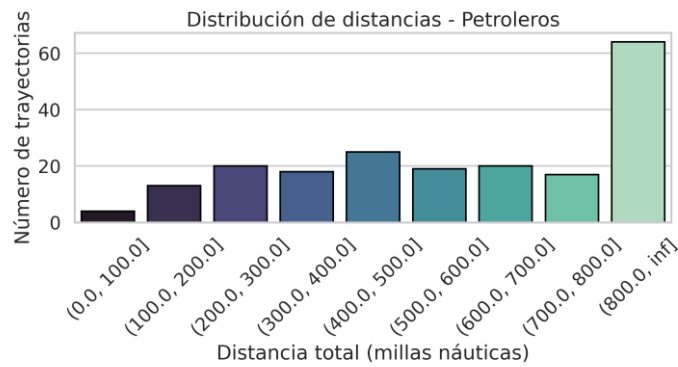


Figura A4-14 Distribución distancias LSTM con filtrado de control para Petroleros

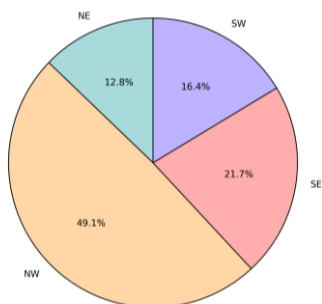


Figura A4-15 Distribución geográfica reales Petroleros

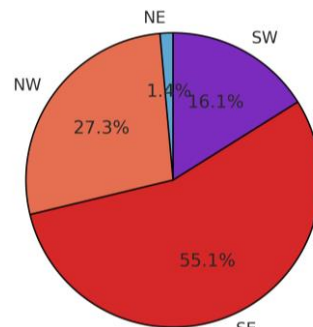


Figura A4-16 Distribución geográfica LSTM con filtrado de control para Petroleros

En términos globales, los resultados muestran que el modelo LSTM reproduce adecuadamente el tipo de trayectoria característico de los petroleros, manteniendo su naturaleza de recorridos de media y larga distancia. No obstante, al igual que en el caso de los cargueros, se observa una tendencia a sobreestimar la duración y la distancia de las trayectorias, lo que indica una ligera desviación en la escala de los recorridos generados.

Pesqueros

A continuación, se analiza el comportamiento del modelo LSTM tras el filtrado para el caso de los buques pesqueros, destacando los aspectos más característicos de este tipo de navegación.

En las **gráficas de velocidad** (Figura A4-17 y Figura A4-18) se mantiene claramente el patrón dominante de bajas velocidades (0–10 nudos), lo que confirma que el modelo conserva uno de los rasgos más distintivos de los pesqueros, asociado a su actividad operativa.

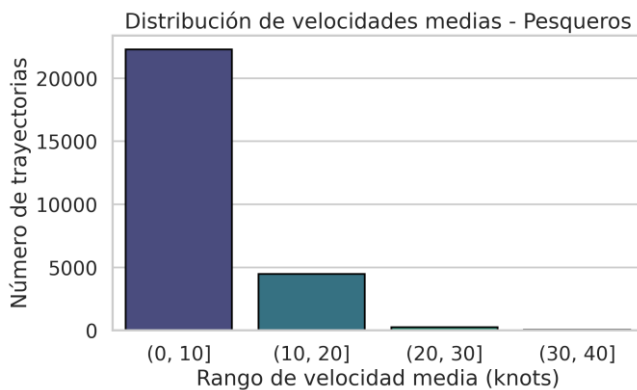


Figura A4-17 Distribución velocidades reales Pesqueros

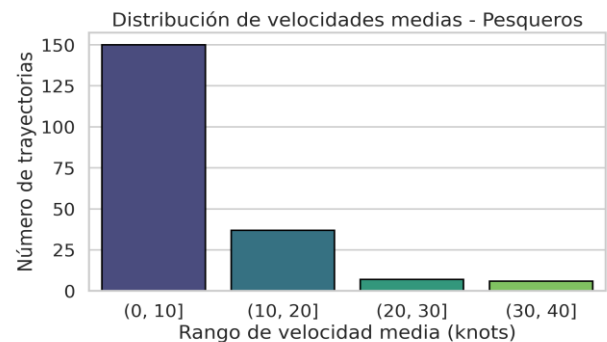


Figura A4-18 Distribución velocidades LSTM con filtrado de control para Pesqueros

En cuanto a la **duración** (Figura A4-19 y Figura A4-20), aunque se mantiene presencia en los intervalos cortos y medios, se observa una mayor dispersión hacia valores superiores, lo que introduce trayectorias más prolongadas de lo habitual para este tipo de buque.

Este efecto se refleja de forma más evidente en la **distribución de distancias** (Figura A4-21 y Figura A4-22), donde, a diferencia de los datos reales —claramente concentrados en distancias cortas—, las trayectorias generadas se desplazan hacia intervalos de distancia media, perdiendo parte del carácter estrictamente local de los pesqueros.

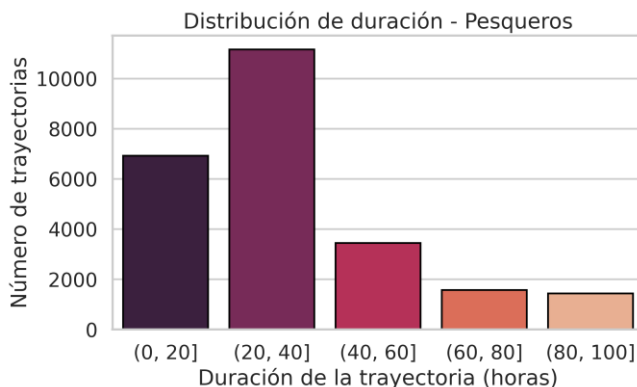


Figura A4-19 Distribución tiempo reales Pesqueros

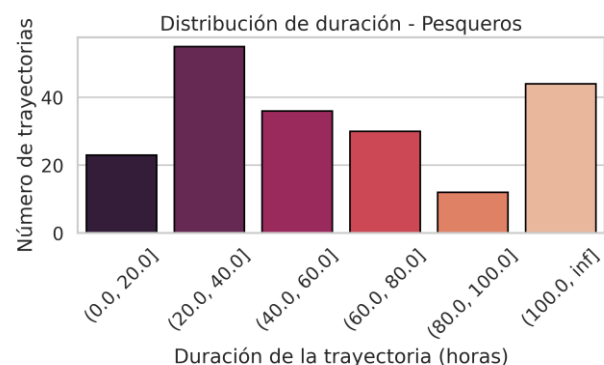


Figura A4-20 Distribución tiempo LSTM con filtrado de control para Pesqueros

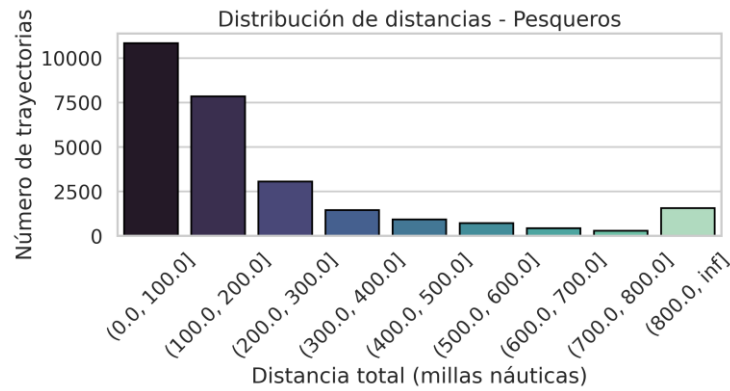


Figura A4-21 Distribución distancias reales Pesqueros

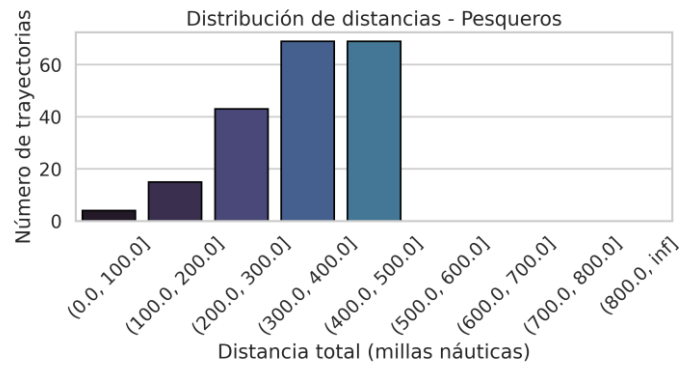


Figura A4-22 Distribución distancias LSTM con filtrado de control para Pesqueros

Por último, en la **distribución geográfica** (Figura A4-23 y Figura A4-24), se observa un cambio significativo en la orientación de las trayectorias, con una fuerte concentración en el cuadrante NE, frente a una distribución más equilibrada en los datos reales. Esto indica una reducción en la diversidad de direcciones de desplazamiento.

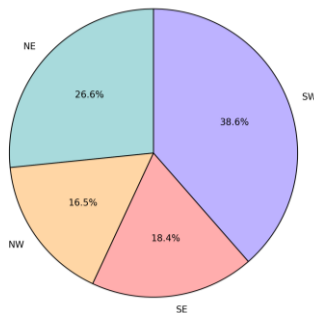


Figura A4-23 Distribución geográfica reales Pesqueros

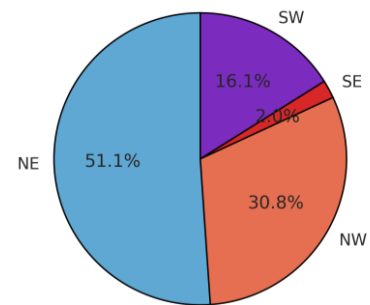


Figura A4-24 Distribución geográfica LSTM con filtrado de control para Pesqueros

En conjunto, las gráficas muestran que, aunque el modelo mantiene correctamente el patrón de baja velocidad característico de los pesqueros, presenta dificultades para reproducir su carácter local, generando trayectorias más extensas y con menor variabilidad direccional.

Pasajeros

A continuación, se analiza el comportamiento del modelo LSTM tras el filtrado para el caso de los buques de pasajeros, caracterizados por trayectorias de corta y media distancia.

En las **gráficas de velocidad** (Figura A4-25 y Figura A4-26) se mantiene la distribución característica, con predominio de los intervalos 0–20 nudos, reflejando adecuadamente la combinación de navegación y maniobra propia de este tipo de buque.

Sin embargo, en las **distribuciones de duración y distancia** (Figura A4-27, Figura A4-28, Figura A4-29 y Figura A4-30) se observa una mayor presencia de trayectorias en intervalos de duración media y larga, lo que se traduce en distancias acumuladas superiores a las observadas en los datos reales. Esto implica una ligera pérdida del carácter regional y relativamente corto de estas rutas.

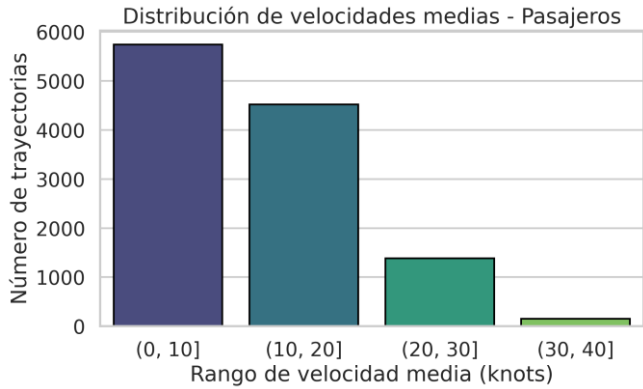


Figura A4-25 Distribución velocidades reales Pasajeros

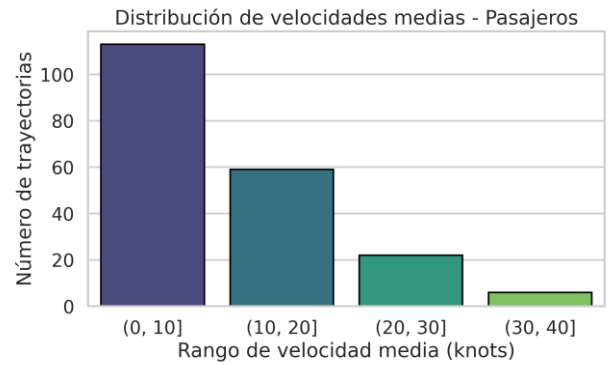


Figura A4-26 Distribución velocidades LSTM con filtrado de control para Pasajeros

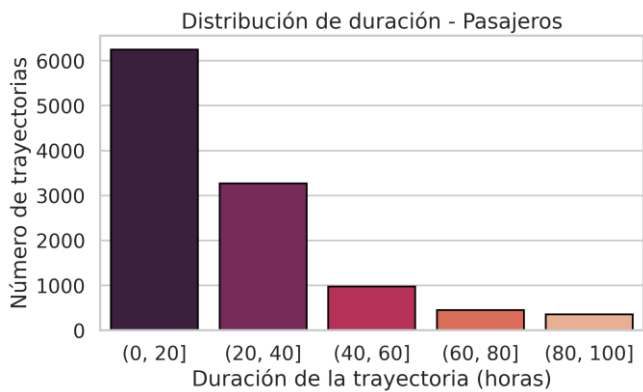


Figura A4-27 Distribución tiempo reales Pasajeros

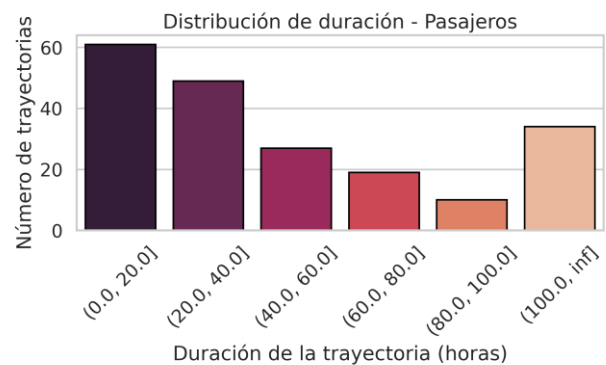


Figura A4-28 Distribución tiempo LSTM con filtrado de control para Pasajeros

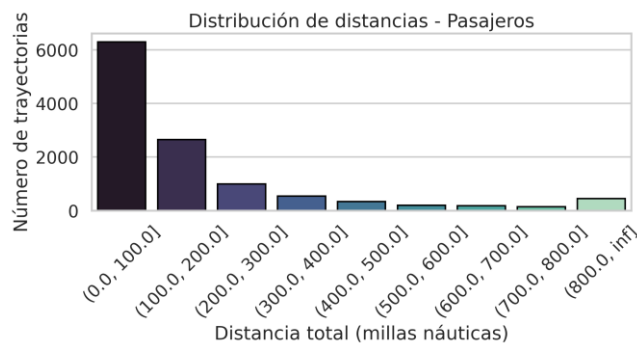


Figura A4-29 Distribución distancia reales Pasajeros

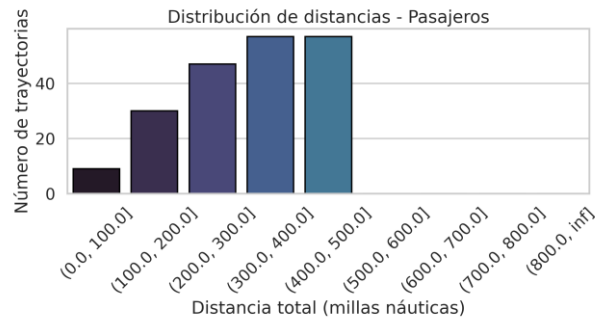


Figura A4-30 Distribución distancia LSTM con filtrado de control para Pasajeros

En cuanto a la distribución geográfica (Figura A4-31 y Figura A4-32), se mantiene la presencia en los distintos cuadrantes, aunque con una mayor concentración en los cuadrantes SE y SW, reduciendo la dispersión respecto a los datos reales.

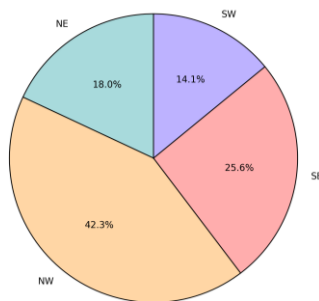


Figura A4-31 Distribución geográfica reales Pasajeros

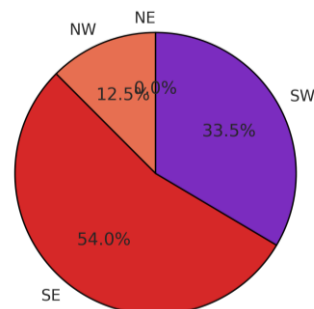


Figura A4-32 Distribución geográfica LSTM con filtrado de control para Pasajeros

En conjunto, las gráficas muestran que el modelo conserva correctamente los rangos de velocidad característicos, pero tiende a generar trayectorias más largas y menos dispersas, alejándose parcialmente del comportamiento regional típico de los buques de pasajeros.

Remolcadores

A continuación, se analiza el comportamiento del modelo LSTM tras el filtrado para el caso de los remolcadores, caracterizados por trayectorias muy cortas y de ámbito local.

En las **gráficas de velocidad** (Figura A4-33 y Figura A4-34) se mantiene claramente la concentración en el intervalo 0–10 nudos, reproduciendo correctamente el patrón de navegación a baja velocidad propio de este tipo de buque.

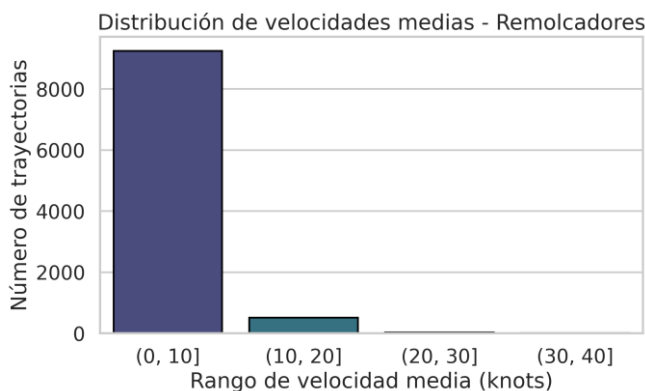


Figura A4-33 Distribución velocidades reales Remolcadores

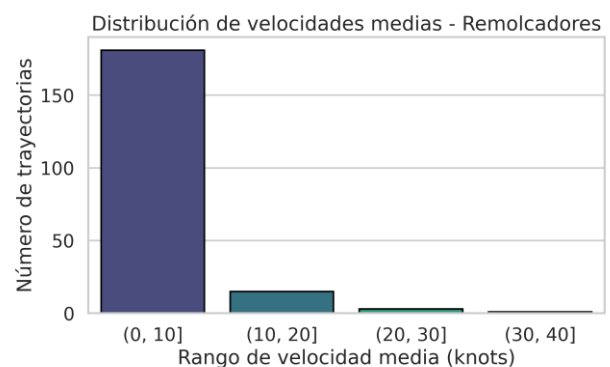


Figura A4-34 Distribución velocidades LSTM con filtrado de control para Remolcadores

Sin embargo, en las **distribuciones de duración y distancia** (Figura A4-35, Figura A4-36, Figura A4-37 y Figura A4-38) se observa una mayor presencia de trayectorias en intervalos superiores a los reales. En particular, las trayectorias generadas se desplazan hacia duraciones medias y distancias más elevadas, perdiendo parte del carácter estrictamente local que define a los remolcadores.

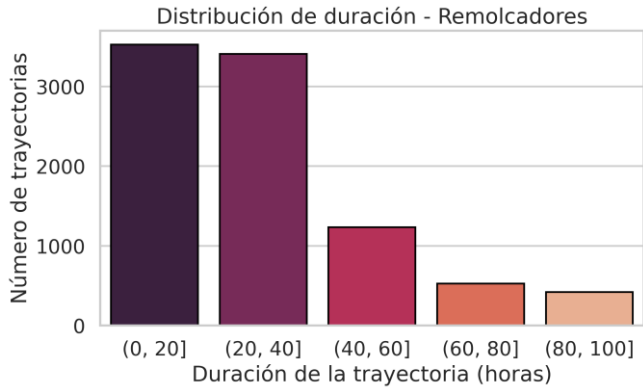


Figura A4-35 Distribución tiempo reales Remolcadores

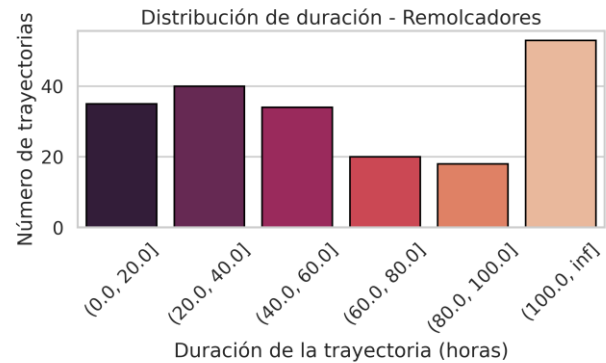


Figura A4-36 Distribución tiempo LSTM con filtrado de control para Remolcadores

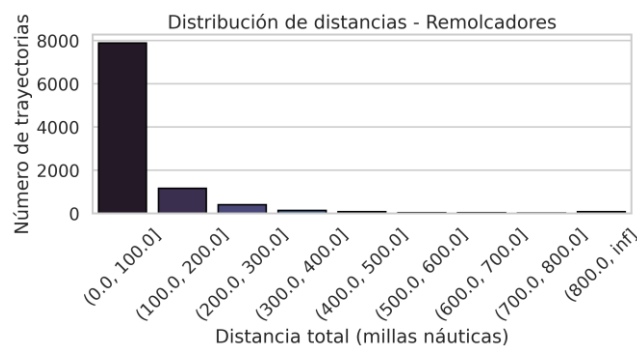


Figura A4-37 Distribución distancia reales Remolcadores

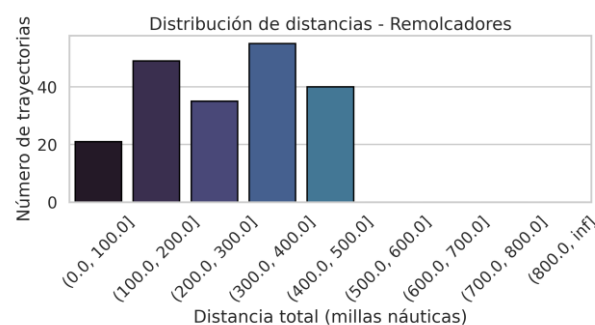


Figura A4-38 Distribución distancias LSTM con filtrado de control para Remolcadores

En cuanto a la **distribución geográfica** (Figura A4-39 y Figura A4-40), se aprecia una fuerte concentración en los cuadrantes SE y NW, frente a una distribución más equilibrada en los datos reales, lo que indica una reducción en la diversidad direccional.

De forma general, las gráficas muestran que el modelo mantiene correctamente el patrón de baja velocidad, pero presenta dificultades para reproducir la escala reducida y el carácter local de las trayectorias, generando recorridos más extensos y menos variados.

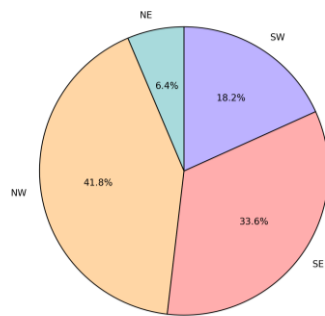


Figura A4-39 Distribución geográfica real de Remolcadores

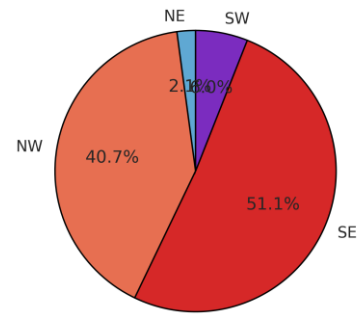


Figura A4-40 Distribución geográfica LSTM con filtrado de control para Remolcadores

En conjunto, el modelo LSTM, tras la aplicación del filtrado, muestra una mejora significativa en la representación de las trayectorias sintéticas, especialmente en lo relativo al control de la escala de los recorridos. El modelo mantiene su capacidad para reproducir los patrones característicos de cada tipo de buque, diferenciando entre trayectorias de largo recorrido y trayectorias de carácter local. No obstante, persisten ciertas limitaciones en la precisión de la duración, la distancia y la dispersión geográfica, lo que indica que, aunque el filtrado corrige parte de las desviaciones iniciales, el modelo aún presenta dificultades para ajustar completamente la variabilidad observada en los datos reales.

Inspección visual de las trayectorias LSTM con filtrado de control

Una vez aplicadas las correcciones al modelo LSTM, se realiza una inspección visual de las trayectorias generadas (Figura A4-41, Figura A4-42, Figura A4-43, Figura A4-44 y Figura A4-45) con el objetivo de evaluar su coherencia espacial y compararlas con los patrones observados en los datos reales.

Siguiendo el criterio establecido en el apartado 4.1, se representan 50 trayectorias sintéticas por tipo de buque, lo que permite analizar los resultados evitando la sobrecarga visual. Esta representación facilita la identificación de patrones generales y posibles desviaciones del modelo.

A nivel global, las trayectorias generadas muestran una mejora respecto a versiones anteriores, especialmente en el control de la escala y en la reducción de comportamientos claramente no realistas. El modelo consigue generar recorridos más coherentes y diferenciados según el tipo de buque, manteniendo la distinción entre trayectorias de largo recorrido y trayectorias de carácter local.

Sin embargo, estas mejoras no eliminan las limitaciones del modelo. La inspección visual pone de manifiesto una reducción de la variabilidad respecto a los datos reales, con trayectorias más homogéneas y menos dispersas. Asimismo, persisten diferencias en la extensión y en la distribución espacial, lo que indica que el modelo no captura completamente la diversidad de patrones de navegación.

Además, se observa una tendencia a generar trayectorias excesivamente suaves, con menor presencia de cambios de rumbo bruscos, lo que reduce el realismo en escenarios donde estas variaciones son habituales. Este comportamiento evidencia una limitación del modelo para reproducir la complejidad dinámica del movimiento real de los buques. Esta característica corresponde al modelo LSTM y no se observa mejora después de la implementación del filtrado de control.

Por tanto, aunque las correcciones introducidas mejoran la coherencia general de las trayectorias, el modelo LSTM sigue presentando dificultades para representar con precisión la variabilidad y riqueza de los datos AIS reales, lo que limita su capacidad como generador completamente realista.

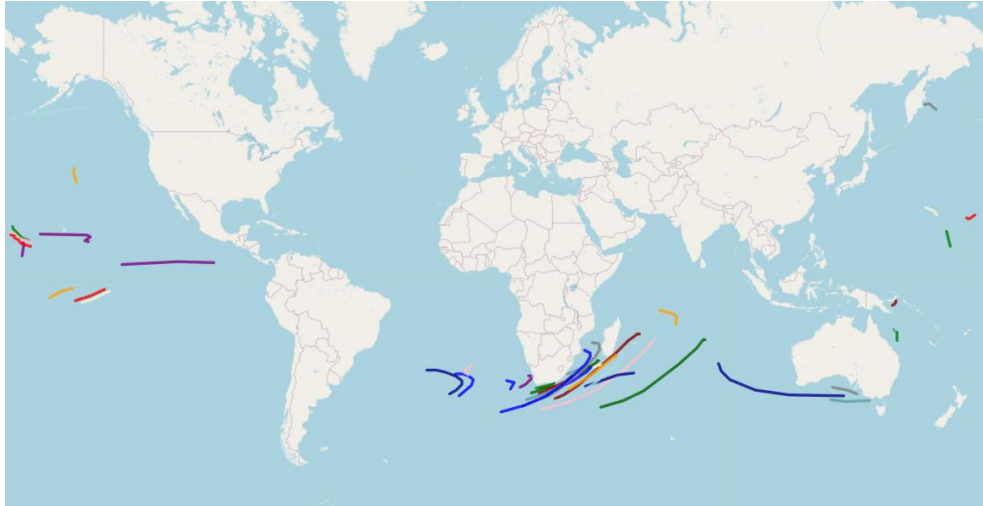


Figura A4-41 Representación de 50 trayectorias sintéticas LSTM con filtrado de control para Cargueros

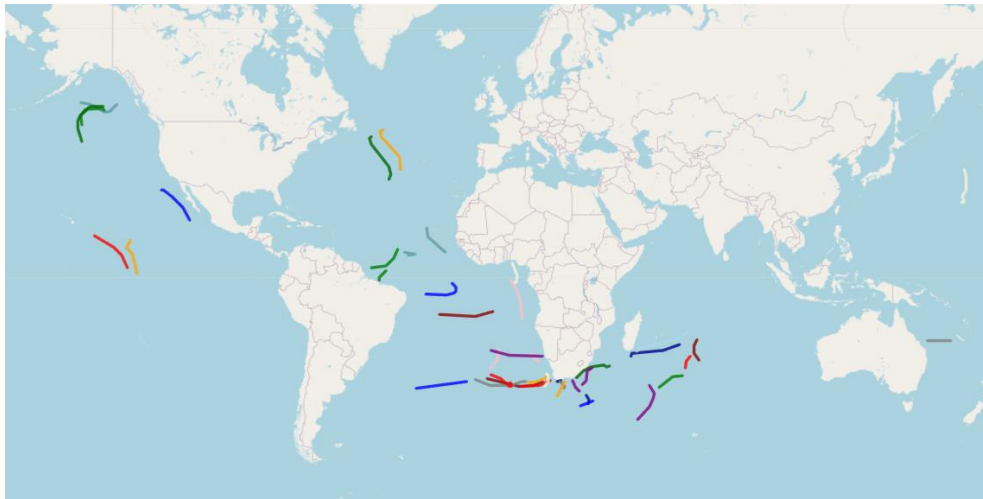


Figura A4-42 Representación de 50 trayectorias sintéticas LSTM con filtrado de control para Petroteros



Figura A4-43 Representación de 50 trayectorias sintéticas LSTM con filtrado de control para Pesqueros



Figura A4-44 Representación de 50 trayectorias sintéticas LSTM con filtrado de control para para Pasajeros



Figura A4-45 Representación de 50 trayectorias sintéticas LSTM con filtrado de control para Remolcadores

ANEXO V: CÓDIGO DE GENERACIÓN DE TRAYECTORIAS CON LSTM CON FILTRADO DE CONTROL Y FILTRADO CONDICIONADO

En este anexo se presenta el código desarrollado para la generación de trayectorias sintéticas de buques mediante un modelo basado en redes neuronales recurrentes del tipo LSTM (*Long Short-Term Memory*). Este código constituye la implementación práctica del modelo descrito en los capítulos anteriores y permite reproducir el proceso completo de generación de trayectorias a partir del modelo ya entrenado con los datos AIS reales.

El procedimiento implementado parte de secuencias reales utilizadas como entrada del modelo, a partir de las cuales se generan nuevas trayectorias de forma iterativa, prediciendo en cada paso las coordenadas geográficas siguientes. En particular, se incluyen dos enfoques de filtrado que permiten evaluar la calidad de las trayectorias generadas, considerando aspectos como la distancia total recorrida y la coherencia espacial de los recorridos.

Asimismo, el código incluye la transformación inversa de las variables previamente normalizadas, permitiendo obtener resultados en unidades geográficas reales. Finalmente, las trayectorias válidas generadas se almacenan en una estructura de datos que facilita su posterior análisis y comparación con los datos reales.

A continuación, se presenta el ejemplo de código correspondiente a cargueros para el **filtrado de control** aplicado al modelo:

```
import geopandas as gpd
from shapely.geometry import Point
import numpy as np
import pandas as pd
import time
from math import radians, sin, cos, sqrt, atan2
from tensorflow.keras.models import load_model

K_POST = 100
N_TRAYS = 200
#Lectura del modelo
model = load_model("modelo_trayectorias_cargueros.keras")

#Calculo de distancia
def haversine_nm(lat1, lon1, lat2, lon2):
    R = 3440.065
    lat1, lon1, lat2, lon2 = map(radians, [lat1, lon1, lat2, lon2])
    dlat = lat2 - lat1
    dlon = lon2 - lon1
    a = sin(dlat / 2)**2 + cos(lat1) * cos(lat2) * sin(dlon / 2)**2
    c = 2 * atan2(sqrt(a), sqrt(1 - a))
    return R * c
```

```
#Evitar que toque tierra
url = "https://naturalearth.s3.amazonaws.com/110m_cultural/ne_110m_admin_0_countries.zip"
world = gpd.read_file(url)
land = world.geometry.union_all()
def trayectoria_toca_tierra(trayectoria):
    longitudes, latitudes = trayectoria
    for lon, lat in zip(longitudes, latitudes):
        punto = Point(lon, lat)
        if land.contains(punto):
            return True
    return False

#Generación de trayectorias
start_time = time.time()
all_rows = []
traj_id = 0
lat_mean, lat_std = limpio_stats["lat"]
lon_mean, lon_std = limpio_stats["lon"]
while traj_id < N_TRAYS:
    idx = np.random.randint(0, len(X_train))
    X_seq = X_train[idx:idx + 1].copy()
    lat_traj = []
    lon_traj = []
    dist_total = 0
    valida = True
    for step in range(K_POST):
        lat_pred, lon_pred = model(X_seq)

        lat_pred = lat_pred.numpy()[0, 0]
        lon_pred = lon_pred.numpy()[0, 0]
        lat_real = lat_pred * lat_std + lat_mean
        lon_real = lon_pred * lon_std + lon_mean
        if len(lat_traj) > 0:
            dist = haversine_nm(
                lat_traj[-1], lon_traj[-1],
                lat_real, lon_real
            )
            dist_total += dist
            if dist_total > 1000: #1000 millas (cargueros)
                valida = False
                break

        lat_traj.append(lat_real)
        lon_traj.append(lon_real)
        new_point = X_seq[0, -1, :].copy()
        new_point[FEATURE_COLS.index("lat")] = lat_pred
        new_point[FEATURE_COLS.index("lon")] = lon_pred

    X_seq = np.concatenate(
        [X_seq[:, 1:, :], new_point.reshape(1, 1, -1)],
        axis=1
    )
```

```

        "traj_id": traj_id,
        "step": step + 1,
        "lat_pred": float(lat_traj[step]),
        "lon_pred": float(lon_traj[step])
    })
    traj_id += 1

df_todas_trajs = pd.DataFrame(all_rows)
end_time = time.time()
print(f"Tiempo total de generación: {(end_time - start_time):.2f} segundos")
print(f"Trayectorias generadas: {traj_id}")

```

```

if valida and not trayectoria_toca_tierra((lat_traj, lon_traj)):
    for step in range(len(lat_traj)):
        all_rows.append({

```

Posteriormente, se incluye el código asociado al **filtrado condicionado**, que introduce restricciones adicionales durante el proceso de validación de las trayectorias generadas.

```

import geopandas as gpd
from shapely.geometry import Point
import numpy as np
import pandas as pd
import time
from math import radians, sin, cos, sqrt, atan2
from tensorflow.keras.models import load_model

K_POST = 100
N_TRAYS = 200
#Lectura del modelo
model = load_model("modelo_trayectorias_cargueros.keras")

#Calculo de distancia
def haversine_nm(lat1, lon1, lat2, lon2):
    R = 3440.065
    lat1, lon1, lat2, lon2 = map(radians, [lat1, lon1, lat2, lon2])
    dlat = lat2 - lat1
    dlon = lon2 - lon1
    a = sin(dlat / 2)**2 + cos(lat1) * cos(lat2) * sin(dlon / 2)**2
    c = 2 * atan2(sqrt(a), sqrt(1 - a))
    return R * c

#Evitar que toque tierra
url = "https://naturalearth.s3.amazonaws.com/110m_cultural/ne_110m_admin_0_countries.zip"
world = gpd.read_file(url)
land = world.geometry.union_all()
def trayectoria_toca_tierra(trayectoria):
    latitudes, longitudes = trayectoria
    for lat, lon in zip(latitudes, longitudes):
        punto = Point(lon, lat)
        if land.contains(punto):
            return True
    return False

```

```
while traj_id < N_TRAYS:
    # 1. Selección de distancia objetivo real
    dist_target = np.random.choice(dist_real_values)
    idx = np.random.randint(0, len(X_train))
    X_seq = X_train[idx:idx+1].copy()
    lat_traj = []
    lon_traj = []
    dist_total = 0
    valida = True

    for step in range(K_POST):
        lat_pred, lon_pred = model(X_seq)
        lat_pred = lat_pred.numpy()[0, 0]
        lon_pred = lon_pred.numpy()[0, 0]
        lat_real = lat_pred * lat_std + lat_mean
        lon_real = lon_pred * lon_std + lon_mean
        if len(lat_traj) > 0:

            dist = haversine_nm(
                lat_traj[-1], lon_traj[-1],
                lat_real, lon_real
            )
            dist_total += dist

        # Rechazo si se excede el umbral máximo
        if dist_total > dist_target * 1.2:
            valida = False
            break

        # Parada si se alcanza la distancia objetivo
        if dist_total >= dist_target * 0.9:
            lat_traj.append(lat_real)
            lon_traj.append(lon_real)
            break

    lat_traj.append(lat_real)
    lon_traj.append(lon_real)

    new_point = X_seq[0, -1, :].copy()
    new_point[FEATURE_COLS.index("lat")] = lat_pred
    new_point[FEATURE_COLS.index("lon")] = lon_pred

    X_seq = np.concatenate(
        [X_seq[:, 1:, :], new_point.reshape(1, 1, -1)],
        axis=1
    )
```

```
if valida and not trayectoria_toca_tierra((lat_traj, lon_traj)):

    for step in range(len(lat_traj)):
        all_rows.append({
            "traj_id": traj_id,
            "step": step + 1,
            "lat_pred": float(lat_traj[step]),
            "lon_pred": float(lon_traj[step])
        })

    traj_id += 1

df_todas_trajs = pd.DataFrame(all_rows)

end_time = time.time()

print(f"Tiempo total de generación: {(end_time - start_time):.2f} segundos")
print(f"Trayectorias generadas: {traj_id}")
```


ANEXO VI: CÓDIGO DE GENERACIÓN DE TRAYECTORIAS CON SEQVAE

En este anexo se presenta el código desarrollado para la generación de trayectorias sintéticas de cargueros (como ejemplo) mediante el modelo SeqVAE. Este código implementa el proceso completo de generación a partir del modelo previamente entrenado, permitiendo obtener nuevas trayectorias que reproducen, en la medida de lo posible, las características de los datos AIS reales.

```
df = pd.read_parquet("trayectorias_cargueros.parquet")
df = df.rename(columns={"id_trayectoria": "trajectory_id"})

# obtener primer punto de cada trayectoria real
df_first_latlon = (
    df.sort_values("timestamp")
      .groupby("trajectory_id", as_index=False)
      .first()[["lat", "lon"]]
)

#Generación
trayectorias = []

for traj, (_, row) in zip(synthetic_real, df_first_latlon.iterrows()):
    lat0 = row["lat"]
    lon0 = row["lon"]

    lat, lon = generar_trayectoria(traj, lat0, lon0)
    trayectorias.append((lat, lon))

rows = []

for traj_id, (latitudes, longitudes) in enumerate(trayectorias):
    for step, (lat, lon) in enumerate(zip(latitudes, longitudes)):
        rows.append({
            "trajectory_id": traj_id,
            "step": step,
            "lat": lat,
            "lon": lon
        })

df_synthetic = pd.DataFrame(rows)
df_synthetic.to_csv("trayectorias_sinteticas_1.csv", index=False)
```