



Departamento de Telemática y Computación (DTC)

TESIS DOCTORAL

Modelado Probabilístico y Análisis de la Demanda en Sistemas de Bicicletas Compartidas: Caso de Estudio BiciMad

Autor: Carlos Miguel Vallez Fernández

Supervisores

Prof. Dr David Contreras Bárcena

Prof. Dr Mario Castro Ponce

Madrid, 2025

***“Todo es hermoso y constante,
Todo es música y razón,
Y todo, como el diamante,
Antes que luz es carbón.”***

*– JOSÉ MARTÍ (1853—1895)
Versos Sencillos 1891*

Dedicatoria

*A mis abuelos Marisa y Fernando,
por su amor, enseñanzas y sacrificios.*

Agradecimientos

Quisiera expresar en este capítulo mi más sincero agradecimiento y gratitud a las numerosas personas y a las instituciones que de una u otra forma han contribuido a que este largo camino se haya podido materializar en el presente documento. Han sido muchas las horas de trabajo invertidas en estos cerca de seis años, muchas las alegrías e incontables los logros parciales conseguidos. Pero también han sido demasiadas las frustraciones y los momentos de angustia y duda. Hoy puedo decir en voz alta que todas y cada una de las experiencias vividas durante la realización de esta tesis han merecido la pena. Y si he llegado hasta aquí os lo debo a vosotros. Sentid esta tesis como parte vuestra porque sin duda todos, de alguna u otra manera, estáis reflejados en ella.

Procedo por tanto a exponer los agradecimientos.

En primer lugar, y como no pudiera ser de otra forma, quiero agradecer a la Universidad Pontificia Comillas en general, y a su dirección la oportunidad que se me dio en su día de poder comenzar a trabajar en la institución. Día a día fui conociendo de la mano de mis compañeros el trabajo del Personal Docente e Investigador y me fui sintiendo atraído por redireccionar mi carrera profesional al ámbito de la docencia e investigación universitaria.

A continuación agradezco a mis dos directores de tesis David y Mario que me hayan acompañado durante este largo proceso, dándome buenos consejos, prestándome su inestimable ayuda incluso en fechas y horas complicadas y en definitiva siendo muy pacientes mientras recorría el camino. Son muchas las cosas a agradecer a ambos pero me gustaría resaltar las que más me han marcado. En el caso de David quiero resaltar mi agradecimiento por ser el inicio de este camino. Por explicarme, cuándo aun no tenía claro empezar con ello, qué era un doctorado y facilitarme todos los trámites iniciales, incluyendo la elección de directores y en definitiva acabar animándome a dar el paso. En lo que respecta a Mario quiero resaltar que se ha convertido para mi en un claro referente a seguir. Agradezco y resalto todas y cada una de sus enseñanzas técnicas y metódicas. Valoro todas las correcciones que me ha realizado. He disfrutado mucho en el camino con los debates relacionados con los lenguajes y entornos de programación, y he de reconocer que al final se salió con la suya y de alguna manera. ^abracé el lado oscuro de la programación incluyendo el lenguaje R en parte de mi código.

Seguidamente quiero acordarme y agradecer a todos mis compañeros de la Universidad, tanto de mi departamento y escuela como de otras facultades, tanto a los que siguen en activo como a los que ya no están en la Universidad y por supuesto tanto a personal docente e investigador como al personal de administración y servicios. Nombrar a cada uno de ellos haría interminable esta sección y seguramente me haría correr el riesgo de dejarme al alguno olvidado. Confío en que, aunque este agradecimiento es general, cada uno lo recibe de manera individual, reconociendo todo lo que le agradezco por su apoyo en esta tesis. A todos, muchas gracias.

No podía olvidarme tampoco de los agradecimientos a todos mis alumnos. Porque aunque hayan aprendido de mi, yo también he aprendido y mucho de ellos.

Agradezco también a mis compañeros y profesores de doctorado destacando especialmente a José Ignacio por su apoyo y ánimos.

Para ir cerrando quiero expresar mi especial agradecimiento a mis familiares destacando a las dos personas más importantes en influyentes. A mi madre Lourdes porque sin su apoyo, sin sus consejos y en definitiva sin ella nada de esto hubiera sido posible. Se lo debo todo. Y a Alejandra, porque has aparecido en el momento preciso brindándome toda tu ayuda y apoyo. Siendo un pilar fundamental y básico para que hoy pueda cerrar con éxito este camino. Cualquier agradecimiento se queda corto.

Para terminar quiero agradecer a cada piedra que me encontré en el camino, que como dice la canción "me enseñó que mi destino era rodar y rodar a cada palo que se atravesó en las ruedas de la bicicleta porque me aportaron la experiencia, la motivación y la fuerza necesarias para levantarme y seguir.

Abstract

Large cities face a set of challenges resulting from accelerated urbanization and the centralization of populations in urban environments. Therefore, issues such as traffic congestion, increased pollution levels, and overburdened transportation systems have intensified. These problems, combined with the need to align with the Sustainable Development Goals (SDGs) in Europe, the rise of the sharing economy, and the concept of Mobility as a Service (MaaS), have driven the emergence of new forms of urban transportation.

In this context, Bike-Sharing Systems (BSS) have emerged as an innovative, popular, and successful solution, becoming a key component in promoting sustainable urban transportation.

The growth of BSS has not only motivated their implementation in an increasing number of cities but also attracted the attention of the scientific community starting in 2010. This initiated a period of contributions and research across various fields of knowledge. This diversity, however, led to terminological fragmentation, complicating the review of BSS-related articles.

To address the issue of terminological fragmentation, this thesis developed a thesaurus to standardize concepts related to BSS. Additionally, a systematic methodology was designed to classify BSS-related articles based on their themes and timelines. This classification allowed for the identification of the rebalancing problem as the primary challenge in this field. Furthermore, it highlighted the need to analyze the historical evolution of these systems, including a detailed review of the BiciMad implementation case. The first part of this thesis, which presents the results of the first article, focuses on these fundamental aspects and lays the groundwork for the rest of the research.

The rebalancing problem consists of two interconnected issues: the first focused on predicting service demand and the second on optimizing the process of distributing bicycles across stations to meet this demand. Addressing both issues requires advanced models to predict and manage the demand for these services.

In the second part of this thesis, presented in the second article, a probabilistic machine learning framework is introduced to model and simulate demand in BSS, with a focus on the BiciMad implementation in Madrid.

The proposed approach uses empirical trip data from 2018 and 2019 to fit and validate theoretical probability distributions, including Gamma distributions for travel distances and Negative Binomial distributions for the number of trips. Unlike traditional regression-based methods, the model described in this thesis incorporates stochasticity and uncertainty, enabling the simulation of alternative scenarios and counterfactual analyses (e.g., "What if...?"). The modeling framework also integrates external variables, such as weather conditions and day types (workdays, weekends, and holidays), to improve accuracy.

Key applications of the model include validating empirical datasets, simulating counterfactual scenarios, and identifying unmet demand due to system limitations, such as the lack of bicycles or docking spaces.

Resumen

Las grandes ciudades se enfrentan a un conjunto de desafíos producto de la acelerada urbanización y la centralización de la población en entornos urbanos a la cual asistimos. Como consecuencia se han intensificado problemas como la congestión del tráfico, el incremento de los niveles de contaminación y la sobrecarga de los sistemas de transporte. Estas problemáticas unidas a la necesidad de alinearse en Europa con los Objetivos de Desarrollo Sostenibles (ODS), al auge de la economía colaborativa y al concepto de movilidad como servicio (*MaaS*, por sus siglas en inglés, “*Mobility as a Service*”) han motivado la aparición de nuevas formas de transporte urbano. En este escenario, los sistemas de bicicletas compartidas (*BSS*, por sus siglas en inglés, “*Bike-Sharing systems*”) se presentan como una solución innovadora, popular y exitosa, convirtiéndose en un componente clave para promover el transporte urbano sostenible.

El crecimiento de los *BSS*, además de motivar su implementación por parte de cada vez más ciudades, atrajo la atención de la comunidad científica a partir de 2010. Comenzó un período de aportaciones e investigaciones en muy diversas áreas de conocimiento. Esta heterogeneidad se tradujo en una fragmentación terminológica que dificultaba la revisión de artículos relacionados con *BSS*.

Para abordar el problema de la fragmentación terminológica, en esta tesis se desarrolló un tesaurus que estandariza los conceptos relacionados con los *BSS*. Adicionalmente, se desarrolló una metodología sistemática que permite clasificar los artículos relacionados con los *BSS* en función de su temática y temporalidad. Esta clasificación permitió identificar el “problema del rebalanceo” como el desafío principal en este ámbito. También, se planteó la necesidad de analizar la evolución histórica de estos sistemas, incluyendo una revisión detallada del caso de implementación de BiciMad. La primera parte de esta tesis, que recoge los resultados del primer artículo, se dedica a estos aspectos fundamentales y sienta las bases necesarias para el desarrollo del resto de la investigación.

El problema del rebalanceo es en realidad la suma de dos problemas: el primero enfocado en la predicción de la demanda del servicio y el segundo centrado en la optimización del proceso de distribuir las bicicletas por las estaciones de tal manera que la demanda anterior quede satisfecha. Aportar solución a ambos problemas requiere, por tanto, de modelos avanzados para predecir y abordar la demanda de estos servicios.

En la segunda parte de esta tesis, representada en el segundo artículo, se presenta un marco de aprendizaje automático probabilístico para modelar y simular la demanda en los *BSS*, centrándose en la implementación de BiciMad en Madrid.

El enfoque propuesto utiliza datos empíricos de viajes correspondientes a los años 2018 y 2019 para ajustar y validar distribuciones teóricas de probabilidad, incluyendo distribuciones Gamma para las distancias de viaje y distribuciones Binomial Negativa para el número de viajes. A diferencia de los métodos tradicionales basados en regresión, el modelo descrito en esta tesis incorpora estocasticidad e incertidumbre, lo que

permite la simulación de escenarios alternativos y análisis contrafactuales (preguntas del tipo *¿qué pasaría si...?*). El modelado también integra variables externas, como las condiciones meteorológicas y el tipo de día (laborales, fines de semana y festivos), para mejorar la precisión.

Entre las aplicaciones clave del modelo destacan la validación de conjuntos de datos empíricos, la simulación de escenarios contrafactuales y la identificación de demandas insatisfechas debido a limitaciones del sistema, como la falta de bicicletas o espacios de anclaje.

Acrónimos ¹

AaS	As a Service
ABC	Artificial Bee Colony
AEMET	Agencia Estatal de Metereología
API	Application Programming Interface
ATS	Adaptive Tabu Search
AWS	Amazon Web Services
BSS	Bike-Sharing Systems
CLT	Central Limit Theorem
CNIG	Centro Nacional de Información Geográfica
CNN	Convolutional Neural Networks
CRO	Chemical Reaction Optimization
CSS	Car-Sharing Systems
CSV	Comma-separated values
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
ECSS	Electric Car-Sharing Systems
EDA	Exploratory Data Analysis
EMT	Empresa municipal de transportes
EABC	Enhanced Artificial Bee Colony
FLM	First and Last Mile problem
GANs	Generative Adversarial Networks
GBMs	Gradient Boosting Machines

¹A pesar de estar escrita en castellano, esta tesis adopta los acrónimos y términos en inglés para ciertos conceptos, con el fin de mantener la nomenclatura predominante en la literatura.

GIS	Geographical Information System
GMMNs	Gaussian Mixture Models
GPS	Global Positioning System
HMM	Hidden Markov Models
ICT	Information and Communications Technologies
ILS	Iterated local search
INE	Instituto Nacional de Estadística
JSON	JavaScript Object Notation
KCA	Keyword Co-occurrence Analysis
KL	Kullback-Leibler
KNN	K-Nearest Neighbors
LSTM	Long Short -Term Memory networks
MaaS	Mobility as a Service
MAP	Maximum A Posteriori
MCMC	Markov Chain Monte Carlo
MILP	Mixed-Integer Linear Programming
ML	Machine Learning
MLE	Maximum Likelihood Estimation
ODS	Objetivos de Desarrollo Sostenibles
PBSS	Public Bike-Sharing Systems
PDP	Pickup and Delivery Problem
PIB	Producto Interior Bruto
RBM s	Restricted Boltzmann Machines
RNN s	Recurrent Neural Networks
RQ	Research Question
SMS	Shared Mobility Systems
SVM s	Support Vector Machines
TSP	Travelling Salesman Problem

VaaS	Vehicle as a Service
VAEs	Variational Autoencoders
VRP	Vehicle Routing Problem
VTC	Vehículo de Transporte con Conductor
VSS	Vehicle-Sharing Systems

Índice general

1. Introducción y motivación	8
1.1. Contextualización	8
1.2. Historia de los BSS	13
1.2.1. Definición y características	13
1.2.2. Historia y generaciones de los BSS	16
1.2.3. Historia de BiciMad	17
1.3. Motivación	20
1.4. Fundamentos del aprendizaje automático	21
1.4.1. Introducción al aprendizaje automático - <i>machine learning</i>	21
1.4.2. Conceptos clave de los modelos deterministas frente a los probabilísticos	23
1.5. Preguntas de investigación y objetivos de la tesis	25
1.5.1. Objetivo general de la tesis	25
1.5.2. Preguntas de investigación - <i>research questions RQ's</i>	25
1.5.3. Objetivos específicos	27
1.6. Esquema de la tesis doctoral	32
2. Estado del arte	34
2.1. Revisión del estado del arte y descripción de la metodología	34
2.2. Identificación de palabras clave	35
2.3. Inventario de artículos de revisión	37
2.4. Fases y áreas clave en la evolución de la investigación en BSS	40
2.5. El problema del rebalanceo	45
2.5.1. Cronología de la revisión de la literatura acerca del problema del rebalanceo	47
2.6. Conclusiones del capítulo	52
3. Análisis exploratorio de los datos (EDA)	54
3.1. Definición del alcance de los datos empleados	54
3.2. Descripción y procedencia de los datasets iniciales	55
3.2.1. Dataset 1 - Viajes realizados	55
3.2.2. Dataset 2 - Información de estaciones de BiciMad	58
3.2.3. Dataset 3 - Información meteorológica	60
3.2.4. Dataset 4 - Calendario de días festivos	61
3.2.5. Dataset 5 - Estado en tiempo real de las estaciones de BiciMad	62
3.3. Extracción y transformación de los datos	64
3.3.1. Proceso de extracción de los datos de viajes	65
3.3.2. Proceso de extracción de los datos del resto de fuentes	65
3.3.3. Procesos de transformación	68

3.3.3.1. Método de cálculo de distancias en los viajes de BiciMad	70
3.4. Análisis exploratorio de variables relevantes	73
3.5. Resultados y conclusiones	75
4. Enfoque de ML probabilístico para el modelado de la demanda en BiciMad	76
4.1. Inferencia de las distribuciones de probabilidad	77
4.2. Modelado de la distancia	78
4.2.1. Validación del ajuste de la distribución mediante test de Kullback-Leibler	80
4.2.2. Ajuste de los datos empíricos a la distribución Gamma	84
4.2.3. Justificación de la idoneidad de la distribución Gamma	85
4.3. Modelado del número de viajes	88
4.3.1. Generación de datos sintéticos y validación interna	91
4.4. Creación de un simulador probabilístico	92
4.5. Descripción de alternativas para el ajuste de distribuciones con enfoque bayesiano	94
4.6. Resultados y conclusiones	97
5. Aplicaciones del Simulador - Modelo	99
5.1. Simulación de la redistribución de viajes variando las estaciones disponibles	99
5.2. Imputación de datos y detección de atribuciones erróneas	103
5.3. Predicción de fallas en la demanda no satisfecha	104
5.4. Resultados y conclusiones	106
6. Conclusiones, contribuciones y futuros desarrollos	108
6.1. Conclusiones de la tesis	108
6.2. Principales contribuciones de esta tesis	110
6.3. Publicaciones y conferencias	111
6.4. Futuras líneas de investigación	112
Apéndices	115
A. Resumen Taxonómico de Algoritmos de Aprendizaje Automático (ML)	116
B. Evolución de los mapas de estaciones de BiciMad	120
C. Comparativa Evolutiva de BiciMad: Periodo 2018-2019 frente al Periodo 2023-2024	127
D. Resumen de los artículos de revisión con sus principales características.	131
E. Resumen de artículos sobre el rebalanceo en los BSS y sus principales características.	137
F. Análisis de viajes diarios según temperatura, tipo de día y precipitaciones	143
G. Parámetros de la distribución Binomial Negativa para viajes por hora según tipo de día y lluvia	146
H. Validación interna de la distribución Binomial Negativa con viajes del 2018	149
I. Función de Popularidad	155

J. Código RStan empleado para la modelización con enfoque bayesiano de las distancias	161
Bibliografía	173

Índice de Figuras

1.1.	MaaS aplicado a una ciudad	9
1.3.	Comparativa de frecuencia de viajes y distancia en función del medio de transporte	10
1.2.	Comparativa de frecuencia de viajes y distancia en función del medio de transporte	11
1.4.	Relación entre la velocidad y la distancia de viaje, diferenciando por modo de transporte	12
1.5.	Mapa mundial de implementaciones BSS	14
1.6.	Resumen Generaciones BSS	17
1.7.	Objetivos de Desarrollo Sostenible (ODS) Agenda 2030	20
1.8.	Esquema Modelos de Demanda de Movilidad	22
1.9.	Representación esquemática de las preguntas de investigación, el objetivo principal, los objetivos secundarios y los capítulos de la tesis.	31
2.1.	Sección del Tesauro	37
2.2.	Evolución temporal de la investigación en BSS por temáticas hasta 2019	41
2.3.	Evolución temporal de la investigación en BSS por temáticas hasta 2019	41
2.4.	Visualización de los conceptos de investigación sobre BSS agrupados	42
2.5.	Visualización del análisis de co-ocurrencia de palabras clave realizado en 2022 y agrupado por temática	43
3.1.	Histograma del número de viajes por distancia para los años 2018, 2019 y 2020	56
3.2.	Estructura de un registro registro del fichero JSON de viajes	58
3.3.	Ejemplos de registros del fichero inventario de las estaciones del servicio BiciMad	60
3.4.	Subconjunto de registros a modo de ejemplo del fichero de valores meteorológico registrados por la estación de medición de Retiro	61
3.5.	Subconjunto de registros del fichero calendario.csv a modo de ejemplo	62
3.6.	Ejemplo de un registro del fichero JSON con información de estaciones de BiciMad en tiempo real	64
3.7.	Ejemplos de registros del fichero inventario de las estaciones del servicio BiciMad	65
3.8.	Ejemplos de salidas del programa que realiza múltiples solicitudes a la API en intervalos de 5 minutos	67
3.9.	Ejemplo de registro guardado en MongoDB	67
3.10.	Ejemplo de registros de información en tiempo real de las estaciones guardado en fichero CSV	68
3.11.	Representación gráfica de cálculo de la distancia basado en una combinación de las distancias Haversine y Mahattan.	72
3.12.	histograma de número de usuarios por rango de edad	74
3.13.	Histograma de número de viajes por hora	75

4.1.	Distribución empírica de viajes entre estaciones a una distancia d para el conjunto de datos de BiciMad de 2018	80
4.2.	Comparación de las distribuciones Gamma por hora y valores de sus parámetros	85
4.3.	Promedio y desviación estándar del número de viajes por hora bajo diferentes condiciones	90
4.4.	Validación visual de la distribución NegBin para días soleados y laborables por hora	91
4.5.	Intervalo (percentiles 2,5 % - 97,5 %) para el número de viajes por hora y tipo de día, obtenido a partir de distribuciones Negbin	92
4.6.	Diagrama esquemáticos de los elementos del simulador	93
5.1.	Problemática de selección de estación alternativa	100
5.2.	Listado de estaciones seleccionadas y sus combinaciones con las distancias correspondientes	101
5.3.	Probabilidad de viaje entre pares de estaciones según la distribución Gamma, normalizada para considerar solo las estaciones seleccionadas	101
5.4.	Recálculo de la probabilidad normalizada eliminado la estación de Santa Cruz de Marcedo y diferencia con respecto a la normalización que la incluye	102
5.5.	Visualización de la redistribución de probabilidades entre las estaciones activas más cercanas a la estación eliminada	103
5.7.	Captura del mensaje de BiciMad del 9 de septiembre de 2021 que corrobora datos erróneos.	105
B.1.	Mapa con la localización e identificación de las estaciones de BiciMad en su inauguración en 2014	121
B.2.	Mapa de estaciones de BiciMad del Contrato Integral de Movilidad Diciembre 2014.	122
B.3.	Mapa estaciones de Bicimad 2015-2018	123
B.4.	Ampliación de estaciones de Bicimad en 2019	124
B.5.	Ampliación de estaciones de Bicimad en 2020	125
B.6.	Ampliación de estaciones de Bicimad en 2023	126
B.7.	Sección del mapa interactivo de estaciones de BiciMad 2024	126
F.1.	Visualización del análisis de viajes diarios según temperatura, tipo de día y precipitaciones	144
F.2.	Regresión Lineal de Viajes Diarios en función de las Variables Meteorológicas y del Tipo de Día	145
I.1.	análisis visual de la popularidad de las estaciones de BiciMad en 2018	156
J.1.	Salida visual de la ejecución del Modelo bayesiano jerárquico para ajustar la distribución Gamma a grupos a nivel de día de la semana	172

Índice de Tablas

1.1.	Estadísticas más relevantes de BSS a nivel global extraídas del informe “ <i>The Meddin Bike-sharing World Map Report</i> ”	15
1.2.	Indicadores clave del Servicio BiciMad Diciembre de 2024	19
1.3.	Clasificación de los algoritmos de Machine Learning por categorías	23
1.4.	Comparación entre enfoques deterministas y probabilísticos en machine learning	24
2.1.	Ejemplos de falta de uniformidad terminológica en la literatura sobre BSS	36
2.2.	Variantes principales empleadas en la literatura para denominar al problema del rebalanceo	46
3.1.	Propiedades cualitativas y cuantitativas de los archivos originales de datos sobre viajes de BiciMad	56
3.2.	Descripción de los campos informados en el fichero JSON original de viajes del servicio Bicimad.	57
3.3.	Ejemplo de una entrada del fichero JSON representada en una tabla.	58
3.4.	Descripción de los campos del fichero inventario de las estaciones del servicio BiciMad.	59
3.5.	Descripción de los campos del dataset meteorológico de la estación Retiro.	61
3.6.	Descripción de los campos del fichero calendario.csv.	62
3.7.	Descripción de los campos del archivo JSON proporcionado por la API de estado en tiempo real de las estaciones de BiciMad.	63
3.8.	Listado y descripción breve de las transformaciones realizadas a los datos originales	70
3.9.	Métodos para calcular la distancia entre dos puntos geográficos	71
3.10.	Distribución de valores, conteos y frecuencias relativas de la variable <i>age_range</i>	74
3.11.	Distribución de valores, conteos y frecuencias relativas de la variable <i>use_typq</i>	74
4.1.	Resumen de las propiedades y de las funciones de probabilidad de las principales distribuciones empleadas en la literatura, para modelar la distancia recorrida en el viaje.	79
4.2.	Contextos que justifican el tratamiento de las distancias como datos continuos en lugar de discretos.	81
4.3.	Descripción y limitaciones de métricas alternativas al test de <i>Kullback-Leiber</i>	82
4.4.	Parámetros y valores de la Divergencia <i>KL</i> para las distribuciones consideradas, y ordenadas de mejor (menor <i>KL</i>) a peor ajuste (mayor <i>KL</i>).	83
4.5.	Listado de parámetros <i>Shape</i> (α , adimensional) y <i>Scale</i> (θ , medido en metros) de las funciones de distribución Gamma obtenidas para cada hora del servicio.	85
4.6.	Subconjunto ejemplo de los parámetros (p y n) de la distribución NegBin para diferentes horas y clasificados por tipo de día y condiciones de precipitación.	90
4.7.	Comparativa del enfoque Frecuentista frente al enfoque Bayesiano para ajustar una distribución probabilística a unos datos empíricos.	97

A.1.	Resumen de los principales algoritmos de aprendizaje automático.	119
C.1.	Tabla comparativa resumen entre BiciMad de 2018 y BiciMad de 2024.	130
D.1.	Resumen de los principales artículos de revisión sobre sistemas de bicicletas compartidas . .	136
E.1.	Resumen de artículos relacionados con los diversos enfoques empleados en la literatura para abordar el problema del rebalanceo de bicicletas en los <i>BSS</i>	142
G.1.	Parámetros de la distribución Negbin por hora según tipo de día y lluvia	148
H.1.	Cálculo de intervalos de confianza y estadísticas clave para datos sintéticos generados con distribuciones NegBin	154
I.1.	Datos de popularidad	160

“Life is like riding a bicycle. To keep your balance, you must keep moving”.

La vida es como montar en bicicleta. Para mantener el equilibrio debes seguir moviéndote.

— Albert Einstein carta a su hijo Eduard, 1930

1

Introducción y motivación

Índice del capítulo

1.1. Contextualización	8
1.2. Historia de los BSS	13
1.3. Motivación	20
1.4. Fundamentos del aprendizaje automático	21
1.5. Preguntas de investigación y objetivos de la tesis	25
1.6. Esquema de la tesis doctoral	32

El capítulo inicial de esta tesis se enfoca en aportar contexto al ámbito de estudio, los sistemas de bicicletas compartidas, el cual es necesario para comprender la problemática que se va a tratar. Adicionalmente, se incluye una breve contextualización del aprendizaje automático, para establecer un marco de conocimiento que permita englobar y sustentar el modelo probabilístico implementado en esta tesis. A continuación, se presenta la justificación de la investigación junto con la enumeración de los objetivos principales de la propia tesis. Finalmente, se proporciona una descripción de la estructura del documento para facilitar su lectura y comprensión. En resumen, este capítulo busca responder a las preguntas: *¿Cuál es el problema de investigación y por qué es relevante su estudio?* *¿Cuáles son los objetivos de la tesis?* *¿Cómo está estructurado el documento de tesis?*

1.1. Contextualización

En un escenario global donde la gestión y uso eficiente de los recursos, y por consiguiente la sostenibilidad, adquieren una creciente relevancia para la sociedad, surge el concepto de *Economía Colaborativa* [Hamari et al., 2016]. Este modelo económico-social tiene como objetivo principal la optimización de los recursos infrautilizados. Promueve que los individuos puedan compartir, intercambiar o alquilar bienes o

servicios que no están plenamente utilizados [Botsman, R. and Rogers, 2010]. A lo largo de las últimas décadas, este modelo ha experimentado un gran crecimiento, de la mano de avances tecnológicos y plataformas digitales. Aplicaciones como BlaBlaCar, Wallapop o AirBnB son ejemplos de esta tendencia.

Simultáneamente y con la misma motivación, surge el modelo “*As a Service*” (*AaS*), derivado de los principios del “*Cloud computing*” [Callegati et al., 2018]. Este enfoque ofrece productos y servicios mediante suscripción, siguiendo el principio de “*pay-per-use*” (consiste en pagar por el uso de un servicio o bien y no por su adquisición). A diferencia de la *Economía Colaborativa* donde los usuarios generalmente son propietarios de los bienes, en el modelo *AaS* la propiedad suele pertenecer a una empresa que es la proveedora del servicio (como se observa en ejemplos como Uber, Microsoft 365 o Amazon Web Services (AWS)).

La integración de ambos conceptos (*Economía Colaborativa* y *AaS*) ha influido notablemente en diversos sectores. En particular, el principio de *AaS* aplicado al ámbito de la movilidad urbana da lugar al concepto de “*Mobility as a Service*” (*MaaS*) [Arias-Molinares and García-Palomares, 2020]. Según [Jittrapirom et al., 2017], *MaaS* es un modelo de distribución de la movilidad que busca unificar distintos medios de transporte —autobuses, trenes, metro, taxis y diversas opciones de movilidad compartida— en un único servicio coordinado, como muestra la imagen 1.1. Normalmente, este modelo se implementa en núcleos urbanos, donde la administración gestiona algunos medios de transporte y coordina el conjunto de los mismos [Maas, 2022].



Figura 1.1: MaaS aplicado a una ciudad. Fuente: Elaboración propia.

El éxito en la implementación de *MaaS* depende de dos aspectos claves: (1) comprender las preferencias de los usuarios a la hora de seleccionar uno u otro medio de transporte, y (2) lograr integrar diversos tipos de transporte.

Refiriéndose al primer aspecto, la literatura ofrece obras como [Moshe E and Lerman, 1985; Ben-Akiva and Michel, 1999; Juan and Willumsen, 2024] que relacionan las preferencias con la *Teoría del umbral de distancia modal* y con *Teoría de elección modal*.

La primera de estas teorías sugiere que los individuos consideran uno u otro medio de transporte en función de un rango de distancia, y superado su umbral deciden optar por otro medio de transporte. La Figura 1.2 representa esta idea con valores promedios. El umbral es propio de cada individuo y se ve afectado

por factores como pueden ser la edad, renta, clima, etc. La propia teoría sugiere que se puede manipular el entorno con el fin de orientar la decisión de elección de transporte hacia opciones más sostenibles [Andersson et al., 2018; Martínez-de Ibarreta et al., 2024].

Por otra parte, la *Teoría de elección modal*, es un marco más amplio que analiza cómo las personas eligen uno u otro medio de transporte, considerando diferentes factores, no solo la distancia, como son el coste, la disponibilidad, la comodidad y la sostenibilidad entre otros.

En la literatura encontramos estudios como el de [Mátrai and Tóth, 2016] que compara la frecuencia de viajes frente a la distancia recorrida en función del tipo de transporte, tal y como se muestra en la Figura 1.3. De forma similar, el estudio de [Varga et al., 2016] ilustra con la Figura 1.4 la relación entre velocidad aparente¹ y distancia de viaje diferenciando entre diversos medios de transporte.

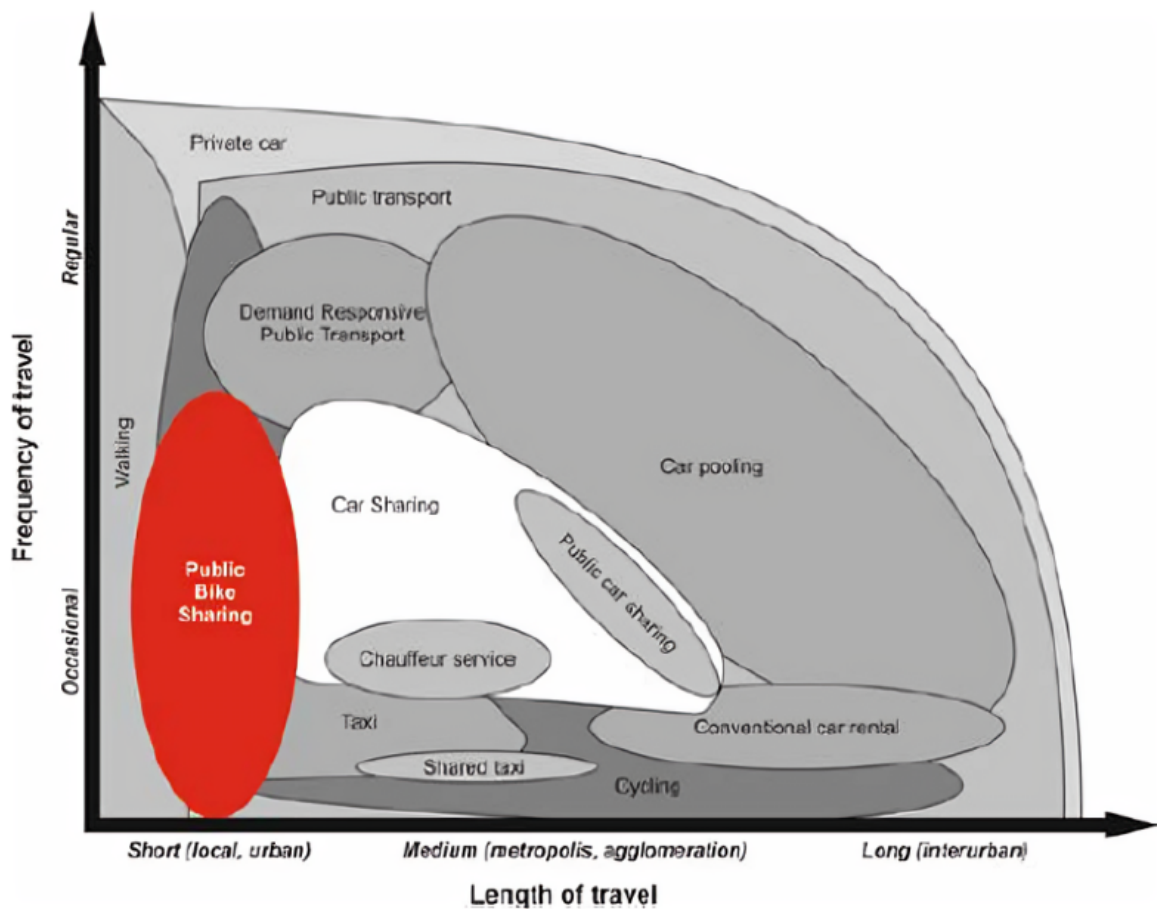


Figura 1.3: Comparativa de frecuencia de viajes y distancia en función del medio de transporte. Fuente: [Mátrai and Tóth, 2016]

¹Velocidad aparente es la Distancia recorrida en línea recta dividida por el Tiempo de viaje.

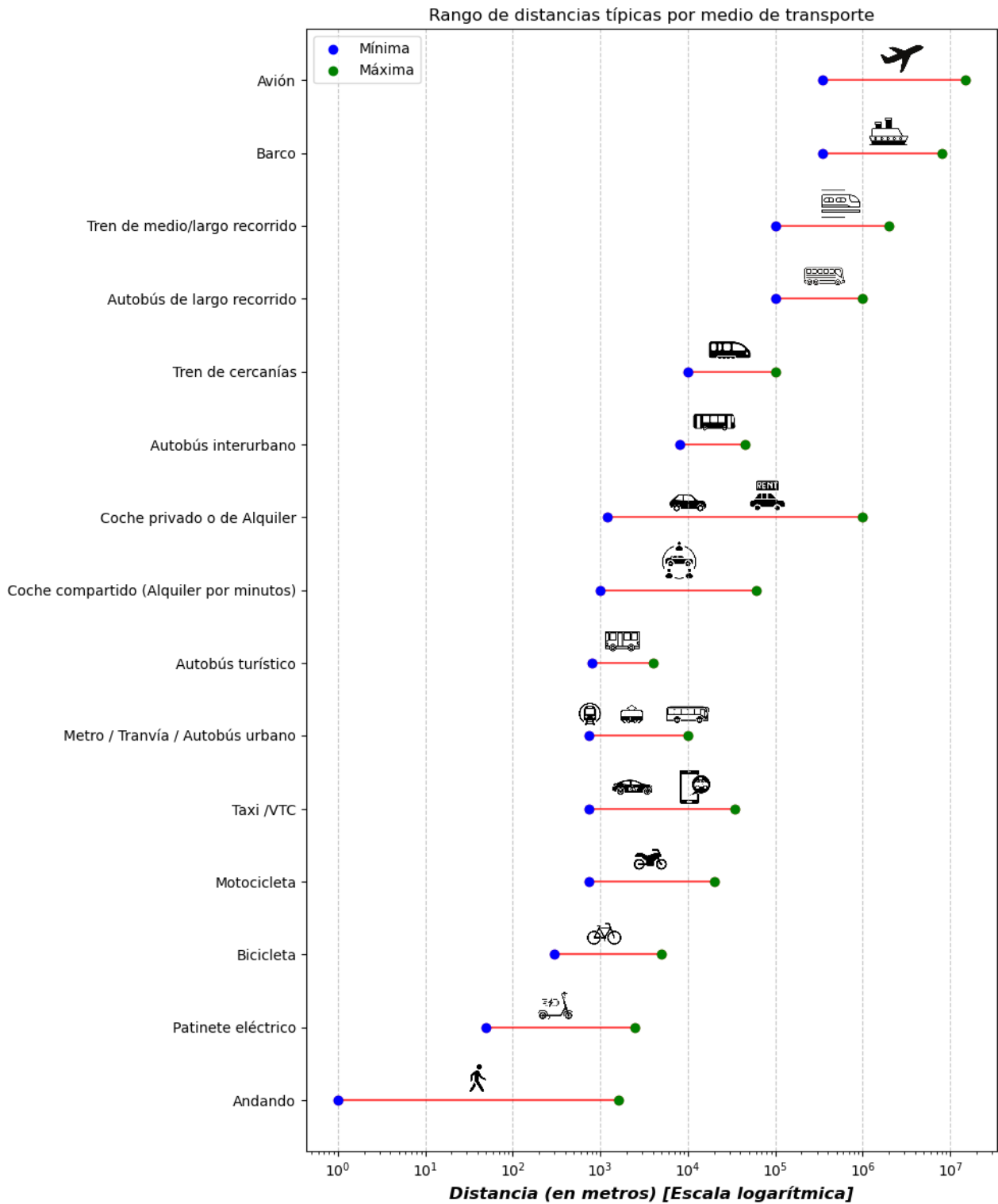


Figura 1.2: Comparativa visual de los rangos típicos de distancias por medio de transporte, ordenadas de menor a mayor. Gráfico de elaboración propia con datos obtenidos de [Eurostat, 2024]

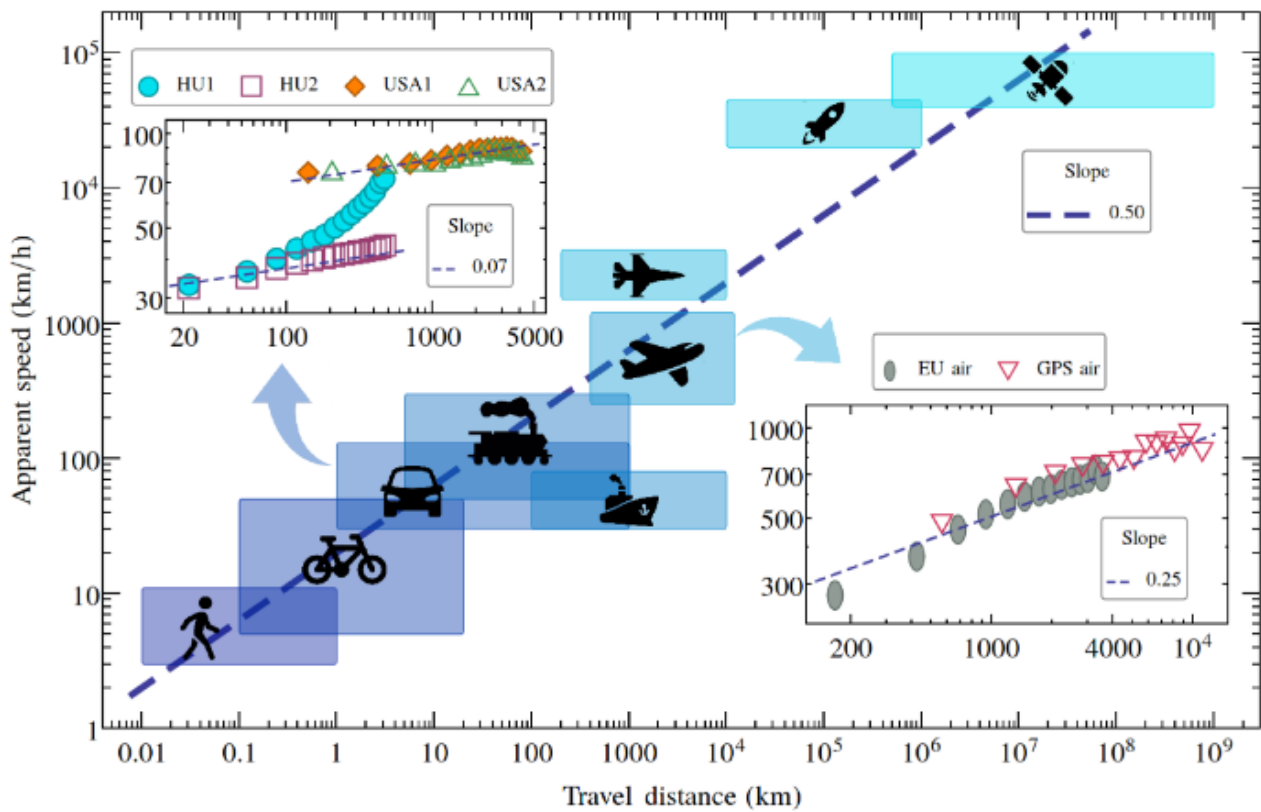


Figura 1.4: Relación entre la velocidad y la distancia de viaje, diferenciando por modo de transporte.

Fuente:[Varga et al., 2016]

Desde la perspectiva de la *Teoría de elección modal*, y con el objetivo de obtener soluciones de transporte más eficientes y sostenibles, dentro del marco de *MaaS* surge el concepto de “*Vehicle as a service*” (*VaaS*), que incluye iniciativas como la movilidad compartida referenciada en la literatura como “*Shared Mobility*” [Santos, 2018]. Esta iniciativa presenta diversas implementaciones ajustadas a los tipos de transporte [Guyader et al., 2021].

“*Car-sharing*” es la primera de ellas, posibilitando el alquiler de un coche por períodos cortos de tiempo y principalmente con el fin de realizar un trayecto específico.

En contraste, y mencionando la opción más reciente, está el “*Scooter-sharing*”. Se trata de los patinetes eléctricos que han hecho aparición en diversas ciudades y permiten sortear las limitaciones que tiene la movilidad en coche, los requisitos regulatorios de la movilidad en motocicleta y las exigencias físicas del uso de bicicletas. Aunque hay que reseñar, en los últimos meses están siendo afectados por estrictas regulaciones derivadas de los problemas que han dado las baterías de ion-litio [Zhang et al., 2018; Jaguemont and Bardé, 2023].

Entre las diversas implementaciones de “*Shared Mobility*”, la opción que emplea las bicicletas como vehículos compartidos (“*Bike-sharing*”) es una de las más exitosas y extendidas [Pucher et al., 2010; Shaheen et al., 2010; Fishman, 2015; Lazarus et al., 2020]. Estos sistemas, referenciados en la literatura científica como “*Bike-Sharing Systems*” (*BSS*), resultan particularmente efectivos abordando el “*problema de la primera y última milla*”² (“*The first and last mile*” (*FLM*) *problem*) del transporte urbano [Dell’Asin et al., 2015; Chiou and Wu, 2024].

²“El problema de la primera y última milla” se refiere al trayecto inicial y final que los viajeros urbanos deben realizar entre su origen y la estación de transporte público o entre dicha estación y su destino final. Las limitaciones en las infraestructuras

Esta tesis se enfoca en el análisis y modelado de la demanda en los *BSS* a partir de datos empíricos del sistema implementado en la ciudad de Madrid denominado BiciMad. El estudio de la demanda, tal y como se detallará en secciones posteriores, no solo es fundamental para entender los patrones de uso del sistema, sino que también resulta primordial para abordar el “*problema del rebalanceo*” (crítico en la operación, sostenibilidad y eficiencia de los *BSS*).

1.2. Historia de los BSS

1.2.1. Definición y características

Los “*Bike-Sharing Systems*” (*BSS*) son una solución práctica y eficiente para el desafío del transporte urbano, permitiendo el uso de bicicletas, distribuidas por diferentes puntos de la ciudad. Mediante estos sistemas es posible tomar una bicicleta en un punto y dejarla en otro, lo que promueve un transporte accesible y asequible, que además contribuye a la reducción del uso de vehículos más contaminantes [DeMaio, 2009]. Idealmente, la oferta y la demanda están homogéneamente distribuidas, y parte del trabajo de redistribución de bicicletas en el municipio se autoorganiza, reduciendo la intervención de los servicios de redistribución. Según el estudio [Kamargianni et al., 2016], que realiza una revisión crítica de los *MaaS*, los *BSS* se han convertido probablemente en la implementación más extendida de *MaaS*. Señala que el éxito de los *BSS* es debido a: (1) La capacidad de poder integrarse con otros modos de transporte como car-sharing, transporte privado o público, (2) El uso de tecnologías de información y comunicación (“*Information and Communications Technologies*” (*ICT*)) para la planificación de rutas, acceso al servicio, disponibilidad de bicicletas, sistemas de pagos, etc. y (3) La promoción de transportes sostenibles frente al uso de vehículos privados. A través de herramientas digitales como [DeMaio, P., 2007; O’Brien, O., 2024], es posible acceder a una visión general de los *BSS* a nivel mundial, así como a información detallada, entre la que se incluyen su estado operativo (en funcionamiento o cerrado), el número total de bicicletas disponibles y la empresa gestora del servicio. En la Figura 1.5 se muestra una captura de una de estas plataformas, mientras que la Tabla 1.1 proporciona un resumen de las estadísticas más relevantes de los *BSS* a nivel global, extraídas del reporte “*The Meddin Bike-sharing World Map Report*” [O’Brien, O. and the Meddin Bike-sharing World Map team, 2022].

urbanas obligan a los viajeros a buscar alternativas como caminar, usar vehículos compartidos como las bicicletas o incluso emplear vehículos privados para cubrir dichas distancias.

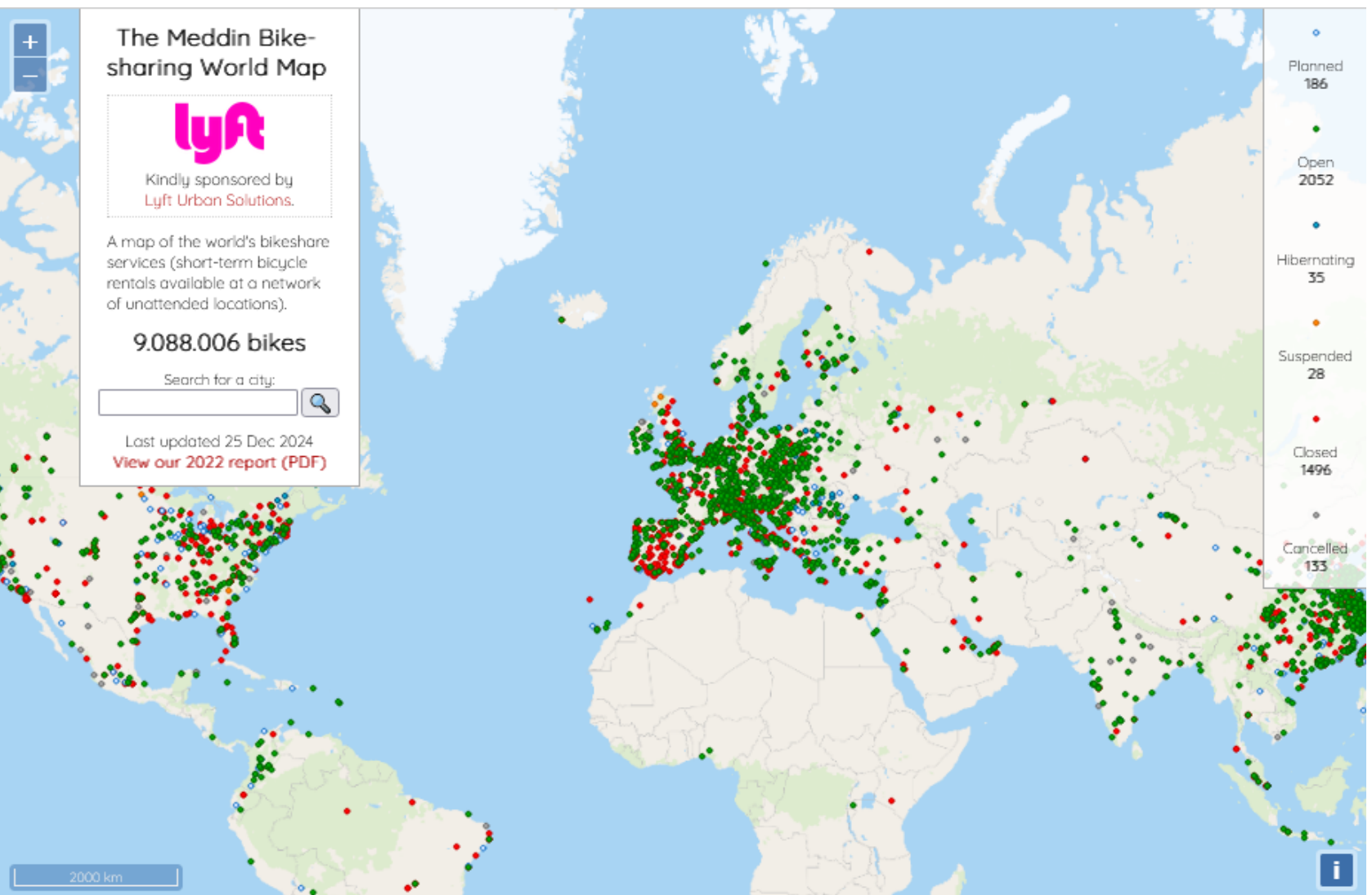


Figura 1.5: Mapa mundial de implementaciones BSS. En verde los servicios en activo y en rojo interrumpidos. Fuente: [DeMaio, P., 2007]

Concepto	Valor
Número de ciudades con sistemas de bike-sharing	1.590
Número de países con sistemas de bike-sharing	92
Continentes con presencia de sistemas de bike-sharing	Todos excepto Antártida
Número de implementaciones de BSS clasificadas por su estado actual (2024)	<i>Activas en funcionamiento:</i> 2066, <i>en planificación:</i> 187, <i>Suspendidas temporalmente o hibernando:</i> 57, <i>Cerradas:</i> 1498 y <i>Canceladas:</i> 133.
Número total de bicicletas en sistemas de bike-sharing	8.967.122
Número de bicicletas con pedaleo asistido (e-bikes)	194.351

Tabla 1.1: Estadísticas más relevantes de BSS a nivel global extraídas del informe “*The Meddin Bike-sharing World Map Report*” [O’Brien, O. and the Meddin Bike-sharing World Map team, 2022]

Aunque las herramientas previamente referenciadas proporcionan una visión global de las distintas implementaciones de BSS, es importante considerar que cada una presenta sus particularidades, las cuales pueden clasificarse según diferentes criterios:

- **Entidad propietaria y gestora del servicio:**
 - **Sistema público:** Propiedad y gestión bajo el control de una Administración o empresa pública.
 - **Sistema privado:** Gestionado y operado por una empresa privada.
 - **Sistema mixto:** Combina la propiedad y gestión entre entidades públicas y privadas.
- **Modelo de operación:**
 - **Estaciones fijas o “Dock-based”:** Si las bicicletas se recogen y se devuelven en puntos específicos denominados estaciones.
 - **Libre disposición, “Free-floating” o “Dockless”:** Si se pueden tomar y dejarlas bicicletas en cualquier parte de la ciudad, respetando las ordenanzas urbanas y normas de circulación).
 - **Híbrida:** Cuando ambos modelos mencionados se ofrecen simultáneamente.
- **Alcance geográfico:**
 - **Con perímetro de uso restringido:** Si el servicio limita el uso de las bicicletas a un recinto determinado, típicamente asociado a una propiedad privada del servicio, como puede ser un campus universitario.
 - **De uso sin limitación:** Cuando el sistema permite el uso sin limitarse a un recinto, pero siempre dentro del área de operación del servicio (zonas de una ciudad, barrios, municipios, etc.).
- **Tecnología utilizada:**
 - **Bicicletas tradicionales:** Sistemas que emplean bicicletas tradicionales que requieren de pedaleo para su funcionamiento.
 - **Bicicletas eléctricas:** Sistemas que incluyen bicicletas con asistencia de pequeños motores eléctricos que facilitan el pedaleo en ciudades con terrenos más desafiantes.

- **Estructura tarifaria:**

- **Gratuito:** Cada vez más difíciles de encontrar.
- **De pago:** Con implementación de muy diversos sistemas y métodos entre los que se encuentran:
 - **Fija con condiciones específicas:** Que permite el uso ilimitado en cuanto a viajes dentro del período contratado y las condiciones acordadas (por ejemplo, viajes no superiores a 30 minutos).
 - **Por tiempo de uso:** El importe depende del tiempo que los usuarios utilizan el servicio.
 - **Suscripción:** Los usuarios contratan un plan por un tiempo (semana laboral, fin de semana, mes, año, etc.) y a cambio obtienen tarifas reducidas o incluso uso limitado por ese período temporal. Pueden estar ligados a la tipología del cliente dando lugar a pases especiales para turistas, por ejemplo.

1.2.2. Historia y generaciones de los BSS

Aunque los *BSS* hayan ganado bastante popularidad en las últimas décadas, lejos de ser un concepto reciente, su origen se remonta a iniciativas de hace muchos años.

Con el fin de facilitar la comprensión de la evolución histórica de los *BSS*, y de acuerdo con la literatura [DeMaio, 2009; Shaheen et al., 2010; Midgley, 2011; Chen et al., 2018; Vallez et al., 2021], detallo las 5 generaciones o fases en las que se divide:

- **Primera Generación:** Se trataba de un sistema rudimentario, caracterizado por la falta de regulación en la toma y devolución de las bicicletas, carácter gratuito del servicio y la identificación de las bicicletas con colores llamativos. Sin incentivos ni penalizaciones por el mantenimiento, el sistema se basaba en la buena conducta y civismo de los usuarios. El proyecto en Ámsterdam (1965) “*Witte Fietsen*” (o “*White Bikes*” en la literatura) fue el pionero, colapsando a los pocos días debido al vandalismo, robos y la confiscación por parte de la policía de las bicicletas abandonadas. Otras iniciativas como “*Vélo Jaunes*” en La Rochelle, Francia (1974) han perdurado en el tiempo con modificaciones en las condiciones, mientras que “*Green bike scheme*” en Cambridge, Reino Unido (1993) fracasó nuevamente debido a robos masivos.
- **Segunda Generación:** La segunda generación de *BSS* mantuvo la diferenciación de bicicletas por color y diseño. Como novedad incorporó el uso de candados y el establecimiento de estaciones de acoplamiento (concepto de “*Dock-based*”), que permitían liberar una bicicleta mediante un pequeño depósito en monedas, el cual se solía reembolsar al devolver y anclarla nuevamente en una estación. A pesar de disponer de una gestión más formalizada, normalmente a cargo de grupos sin ánimo de lucro, se mantenían problemas de robo debido al anonimato de los usuarios y a la ausencia de restricciones temporales en el uso. Los primeros *BSS* de esta generación aparecieron en Dinamarca (Farsø y Grenå en 1991 y Naksø en 1993), destacando el de *Bycyklen* o *City bikes* como el primero a gran escala en Copenhague (1995). A este último le siguieron otras implementaciones en Europa.
- **Tercera Generación:** Supuso la modernización de los *BSS* mediante la incorporación de avances tecnológicos en diversos ámbitos del sistema: anclajes y candados electrónicos, sistemas de telecomunicaciones como GPS, nuevos sistemas de pago como tarjetas inteligentes, identificación de agentes involucrados en el servicio (usuarios, bicicletas, viajes, estaciones, etc.) y accesos por teléfonos móviles entre otros. Estas innovaciones permitieron mejorar la gestión y optimización del servicio, facilitando

el seguimiento de los alquileres y acelerando su expansión internacional. “*Bikeabout*” de la Universidad de Portsmouth en Inglaterra (1996) inició esta generación permitiendo el uso de las bicicletas a los estudiantes mediante tarjetas magnéticas. Seguidamente aparecieron “*Vélo à la Carte*” en Rennes, Francia (1998) y “*Call a Bike*” en Múnich, Alemania (2000).

- **Cuarta Generación:** Se caracteriza por la incorporación de mejoras al servicio como el uso de paneles solares para alimentar las estaciones y la introducción de bicicletas eléctricas. Se realizaron mejoras en el interfaz de usuario de acceso al servicio. Además destaca la integración con otros medios de transporte, por ejemplo con tarjetas de transporte multimodal. Sin embargo lo más relevante es el uso de la información disponible para optimizar la distribución de bicicletas, moviéndolas desde estaciones con alta ocupación a estaciones con alta demanda. ejemplos clave de esta generación son “*Velo’v*” en Lyon, Francia (2005) y “*Vélib*” en París, Francia (2007). El servicio BiciMad de Madrid forma parte de esta generación.
- **Quinta Generación:** La característica diferenciadora de esta generación es la supresión de la necesidad de las estaciones tradicionales (aparece el concepto de “*Free-floating*” o “*Dockless*”). La bicicletas se pueden dejar en cualquier lugar y su localización para ser usadas se realiza mediante aplicaciones móviles. Además estos sistemas incorporan el uso de tecnologías Big Data para procesar los datos del servicio. Entre los primeros ejemplos de esta generación destacan los BSS “*Ofo*” de Beijing (2015) y “*Mobike*” de Shanghai (2016) en China.

Después de haber revisado las 5 generaciones de BSS, el gráfico a continuación muestra una línea temporal resumida en la Figura 1.6.

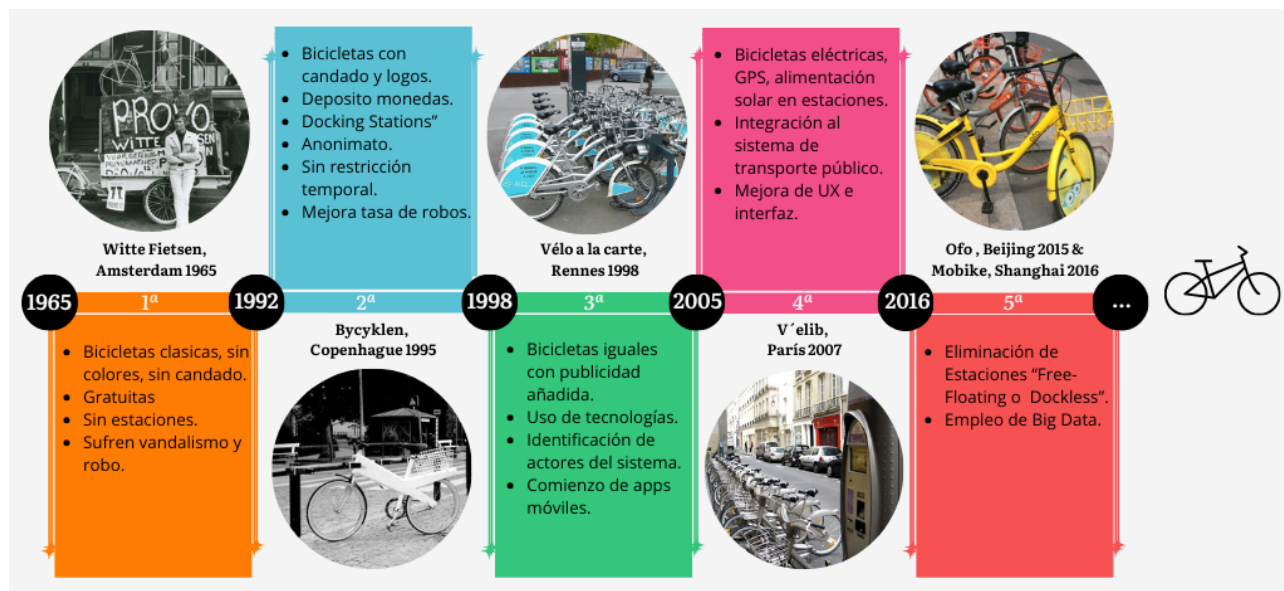


Figura 1.6: Resumen Generaciones BSS. Fuente: Elaboración propia.

1.2.3. Historia de BiciMad

Para comprender el surgimiento de BiciMad, resulta esencial analizar las características de Madrid, su entorno urbano y su sistema de transportes.

En calidad de capital de España, Madrid se presenta como el área urbana con mayor densidad poblacional del país, al contar con 3,2 millones de habitantes en la ciudad y entorno a los 6 millones en el área metropolitana en 2018, el año seleccionado para el presente estudio, de acuerdo con los datos publicados por el Instituto Nacional de Estadística (INE) [INE (Instituto Nacional de Estadística, 2024)]. Su extensión es de 604,45 Km², resultando en una densidad de población aproximada de 5300 hab/Km².

Respecto a la orografía, Madrid se caracteriza por un relieve variable, con una altitud media de 657 metros sobre el nivel del mar. Sus elevaciones oscilan entre los 543 metros, en la cuenca del río Manzanares, y los 846 metros, en el Monte de El Pardo [Centro Nacional de Información Geográfica (CNIG), 2024].

Climatológicamente, la ciudad presenta marcados contrastes estacionales, con veranos calurosos que fácilmente superan los 30°C e inviernos fríos que se acercan a los 0°C. Las precipitaciones son moderadas e irregulares, con mayor frecuencia en otoño y primavera, mientras que los veranos son mayormente secos [Instituto de Estadística (IE) de la Comunidad de Madrid, 2024].

En el ámbito económico destaca que el Producto Interior Bruto (PIB) per cápita era de 41600 € en 2018, siendo el 12 % del PIB nacional (19 % si se incluye toda el área metropolitana) [Varela, A F - Business Insider, 2019].

Finalmente, y en lo que a sistema de transportes respecta, Madrid presenta un sistema integral, incluyendo servicios públicos y privados. En las áreas metropolitanas el vehículo privado representa el 57,8 % de los desplazamientos, mientras que en Madrid ciudad el transporte colectivo y los modos no motorizados superan el 37 %, relegando al vehículo privado al 24,7 % según [Vassallo et al., 2021]. Madrid cuenta con una red de metro de 295 km, con 13 líneas y 302 estaciones Metro Madrid [Metro de Madrid S.A, 2023]. Adicionalmente dispone de un servicio de autobuses urbanos operados por la EMT con 219 líneas y 10.790 paradas [Empresa Municipal de Transportes de Madrid S. A. (EMT Madrid), 2024]. A estos dos medios de transportes hay que añadir un servicio de taxis municipal, varias empresas de servicio de VTC, y numerosas opciones de servicios de vehículos compartidos como coches, motocicletas, patinetes y bicicletas, destacando el servicio de BiciMad.

Habiendo contextualizado las características fundamentales de la ciudad de Madrid, se pasa a analizar la historia y evolución del sistema de bicicletas públicas BiciMad.

El servicio de BiciMad fue inaugurado el 23 de junio de 2014, dando cobertura a seis distritos de Madrid con 1,560 bicicletas y 123 estaciones. Los inicios del servicio se caracterizaron por una fuerte promoción del transporte sostenible y problemas de vandalismo, que afectaron a la flota de bicicletas. Estos desafíos operativos crearon la necesidad de tener una mayor inversión en seguridad y mantenimiento, áreas que no habían sido suficientemente previstas en la planificación inicial del servicio. Por ello, en 2016, y tras varias incidencias en el sistema que no pudieron ser gestionadas por la empresa concesionaria inicial Bonopark, el Ayuntamiento de Madrid decide municipalizarlo.

La incorporación de BiciMad al ámbito municipal fue clave para aumentar la eficiencia, reducir quejas y alcanzar un uso récord en 2017. Todo esto se materializó en los años siguientes con la extensión del servicio más allá de los límites de la M-30 (Almendra central) introduciendo nuevas estaciones, la adquisición de más bicicletas y un incremento del número de usuarios del servicio. Adicionalmente, la administración fortaleció la infraestructura técnica, minimizando las interrupciones por averías y el vandalismo, lo que permitió un servicio más seguro y eficiente.

Como consecuencia de la pandemia de COVID-19, el servicio estuvo inactivo del 16 de marzo al 27 de abril de 2020. Pese a estas circunstancias, el 19 de junio se registraron 17.338 viajes en el servicio, marcando un récord de viajes en un solo día. En agosto de 2020, el Ayuntamiento incorporó BiciMad Go como una alternativa complementaria al sistema original BiciMad. Este nuevo servicio, con un coste más elevado, permitía a los usuarios tomar y dejar bicicletas en cualquier lugar de la ciudad sin ser necesariamente una estación de BiciMad. Esto se tradujo en una mayor libertad de movimiento, aumentando la universalidad

del servicio y llegando a zonas de la ciudad sin disponibilidad de estaciones convencionales. Inicialmente, esta extensión del servicio contó con 454 bicicletas.

A comienzos del año 2021, una nueva interrupción del servicio fue causada por la borrasca Filomena, responsable de una de las mayores nevadas en la capital el 8 de enero. Este hecho no impidió que durante ese mismo año el servicio se extendiera a 15 distritos.

En 2022, el Ayuntamiento anuncia una inversión 48,8 millones de euros en el servicio, de los cuales 30 millones procedían de fondos europeos *Next Generation EU*, con el fin de extender el servicio a todos los distritos de la ciudad y modernizarlo. Al final el proyecto se tradujo en una renovación completa del servicio, la cual incluía una nueva aplicación para móviles, renovación de las bicicletas, reforma de estaciones, presentación de un nuevo sistema tarifario y un cambio estratégico en la imagen y posicionamiento de la marca. Este proceso se llevó a cabo durante un periodo aproximado de un año, en el cual el sistema anterior y el nuevo coexistieron. Durante el período de convivencia el servicio permaneció gratuito inicialmente hasta julio de 2023, pero con posteriores y sucesivas extensiones hasta diciembre de 2023 y enero de 2024, a pesar de que el 26 de mayo de 2023 se había finalizado el proceso de substitución estaciones antiguas por las nuevas.

El 17 de noviembre de 2023, dentro del período de gratuidad del servicio, BiciMad batió su récord histórico de desplazamientos en un solo día, alcanzando la cifra de 57.116 desplazamientos únicos.

Todas las fases descritas se encuentran detalladas en el Apéndice B, donde se incluyen diversas figuras que ilustran las ampliaciones del servicio a lo largo del tiempo.

Finalmente, en 2024 se producen eventos que tienen un impacto directo en el servicio. Por un lado, en octubre de 2024 tiene lugar la última ampliación de estaciones registrada hasta la fecha, permitiendo la llegada del servicio a áreas previamente no cubiertas. Por otro lado, el Ayuntamiento revoca las licencias de operación a competidores directos de BiciMad, como otras empresas de bicicletas compartidas y compañías de patinetes eléctricos.

La tabla 1.2 resumen algunos valores estadísticos de BiciMad a fecha de diciembre de 2024.

Concepto Estadístico	Valor
Número de usuarios BiciMad	136.081 suscriptores
Número de usos mensuales BiciMad	419.368 viajes
Número de usos acumulados año 2024	9.542.223 viajes
Tiempo medio de viaje con bicicleta	15,49 minutos
Distancia media recorrida	2.661 metros
Número de estaciones	630 estaciones
Número de bicicletas disponibles en el servicio	7.735 bicicletas
Media de usos por bicicleta y día	3,75 usos por bicicleta y día

Tabla 1.2: Indicadores clave del Servicio BiciMad Diciembre de 2024

En resumen, desde su inauguración en 2014, el servicio de BiciMad ha pasado por múltiples etapas de transformación, influenciadas por los cambios en la administración municipal. Mientras que los años 2018 y 2019 pueden considerarse un periodo de estabilidad en su funcionamiento, los años posteriores se vieron afectados tanto por eventos externos, como la pandemia de COVID-19 y la borrasca Filomena, como por modificaciones estructurales en el propio servicio. Estas circunstancias han provocado cambios significativos tanto en la demanda del servicio como en el comportamiento de los usuarios. Se requiere por tanto, de un periodo de adaptación para que el sistema se estabilice. Es entonces cuando se podrá aplicar el modelo descrito en esta tesis, adaptando su parametrización a las nuevas condiciones.

Para entender los cambios que ha sufrido BiciMad desde 2018, se presenta la Tabla C.1 comparativa en el apéndice C, que muestra diferencias clave en aspectos como el número de estaciones, bicicletas y diseño del

sistema. Estos cambios justifican que el análisis de esta tesis se enfoque en el servicio durante 2018 y 2019. Una vez el sistema actual se establezca, el modelo desarrollado podrá aplicarse, ajustando su parametrización para adaptarse a las nuevas condiciones.

1.3. Motivación

En 2015 los Estados Miembros de la *Organización de las Naciones Unidas (ONU)* establecieron una hoja de ruta global para abordar los desafíos del siglo XXI, denominada *Agenda 2030* [Krehbiel et al., 2017]. La agenda incluye 17 *Objetivos de Desarrollo Sostenible (ODS)* nombrados en la Figura 1.7, que abordan cuestiones como la pobreza, la salud, la acción climática, etc. con un enfoque holístico y global. Los ODS establecen metas específicas que deben alcanzarse para el 2030. De entre los 17 ODS son de interés para esta tesis: (1) ODS 3 “Salud y bienestar”, (2) ODS 7 “Energía Asequible y No Contaminante”, (3) ODS 9 “Industria, Innovación e Infraestructura”, (4) ODS 11 “Ciudades y Comunidades Sostenibles” y (5) ODS 13 “Acción por el Clima”. Al analizar las características y el comportamiento del sistema BiciMad, contribuyen a los ODS 11 y 13 promoviendo el desarrollo de sistemas de transporte sostenibles y la mitigación del impacto ambiental. Además, la simulación de la demanda de BiciMAD mediante modelos probabilísticos están estrechamente relacionadas con los ODS 9 y 7 ya que promueven soluciones innovadoras orientadas a optimizar la eficiencia energética y minimizar las emisiones de carbono. Finalmente resaltar que, de manera transversal, las preguntas de investigación planteadas en el apartado 1.5.2 contribuyen al ODS 3 al fomentar el uso de transporte activo, como las bicicletas compartidas, lo que impacta positivamente en la salud de los usuarios y en la calidad del aire.



Figura 1.7: Objetivos de Desarrollo Sostenible (ODS) Agenda 2030 ONU.

Fuente: [Naciones Unidas, 2015]

En este contexto, los “*Bike-Sharing Systems*” (BSS) se presentan como una solución fundamental, ofre-

ciendo un transporte eficiente, sostenible (con bajas emisiones) y alineado con los Objetivos de Desarrollo Sostenible (ODS).

La investigación en torno a los BSS ha experimentado un notable crecimiento, abarcando múltiples disciplinas y campos del conocimiento, como detallamos en el Capítulo 2. El interés académico se centra en gran medida en los patrones de uso del servicio, cuya comprensión y análisis es fundamental para abordar el “Problema del Rebalanceo”, ampliamente discutido en la literatura. Este problema inicialmente surge de la “demanda desbalanceada y asimétrica” que caracteriza a los BSS basados en estaciones “*Dock-based*”. En estos sistemas algunas estaciones se pueden quedar sin bicicletas por alta demanda, mientras otras carecen de espacio para dejarlas por alta ocupación de los anclajes. Para mitigar este problema se emplean vehículos que redistribuyen las bicicletas entre las estaciones. En los sistemas “*Free-floating*” el problema es análogo aunque más complejo al no contar con puntos fijos para la entrega y recogida de bicicletas.

El “Problema del Rebalanceo” se divide en realidad en dos problemas: (1) “El problema del modelado de la demanda” y (2) “El problema de la optimización de la redistribución”. En esta tesis nos centraremos en el primer problema, ya que el segundo problema es un clásico de la literatura abordado en múltiples áreas de conocimiento y con diversas técnicas como se recoge en [Caceres-Cruz et al., 2014; Golden et al., 2023; Liu et al., 2023]. La clasificación de los Modelos de Demanda en Movilidad según [Barbosa et al., 2018] se divide en dos categorías principales: (1) Modelos a Nivel Individual que analicen la casuística de la movilidad basándose en características particulares de cada individuo o viaje; (2) Modelos a Nivel de Población que estudian la movilidad de forma agregada, basándose en matrices Origen-Destino (Matrices O-D) y el Modelo Gravitacional. Este modelo se inspira en la “Ley Gravitacional de Newton” asumiendo que la movilidad entre dos lugares está relacionada con su nivel de “atracción”(la cual a su vez depende de factores como población, localización geográfica y orografía, economía, etc). Esta tipología de modelo adolece de problemas como: (1) Sobreajuste debido a la “masa” de la estación destino, mala generalización para datos nuevos y la no captura de la aleatoriedad.

Todo esto ha motivado que en esta tesis se investigue y desarrolle una variante de los Modelos de Demanda en Movilidad a Nivel de Población, con un enfoque y método de Machine Learning Probabilístico. El cual, además de solventar algunos problemas mencionados del “Modelo Gravitacional”, incluye la distribución de los propios datos (explicándolos con precisión y mejor que las matrices O-D) y permitiendo la generación de datos sintéticos.

El esquema de la Figura 1.8 muestra la clasificación de los Modelos de Demanda de Movilidad, con algunas de sus características e indicando exactamente dónde se sitúa el aporte de esta tesis.

1.4. Fundamentos del aprendizaje automático

La construcción de un modelo probabilístico, como el que se plantea en esta tesis, requiere un entendimiento previo de los conceptos básicos del aprendizaje automático. Esta sección se centra en explorar y contextualizar de manera breve y concisa los fundamentos de esta disciplina, destacando sus principales enfoques y clasificaciones. Adicionalmente, se incorporará una selección de referencias de libros reconocidos en el campo del aprendizaje automático, con el propósito de ofrecer recursos complementarios si se desea profundizar en los conceptos abordados en esta sección.

1.4.1. Introducción al aprendizaje automático -*machine learning*

El aprendizaje automático (*ML*) es una rama de la inteligencia artificial (*IA*), que se centra en el diseño de algoritmos para construir modelos capaces de aprender y analizar datos para identificar patrones y tendencias de forma autónoma. Su objetivo principal es el de posibilitar que los ordenadores realicen estas ta-

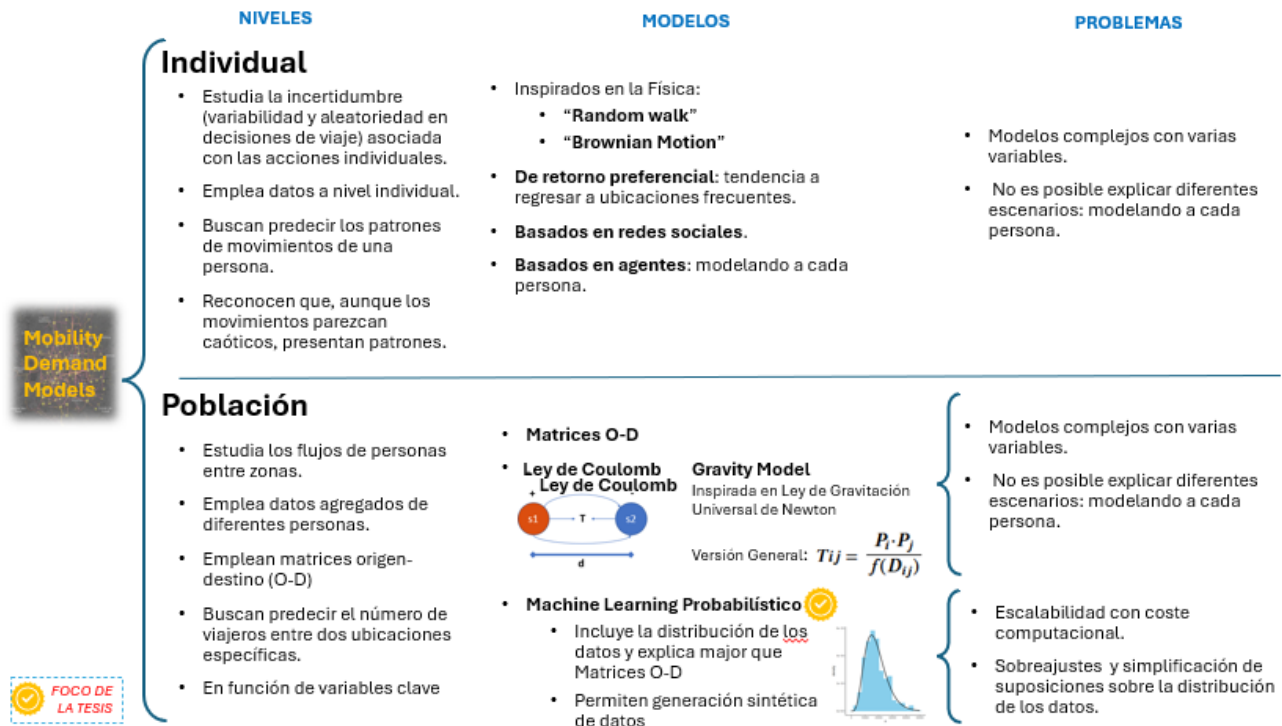


Figura 1.8: Esquema Taxonomía de Modelos de Demanda de Movilidad con localización del aporte de esta Tesis. Fuente: Elaboración propia.

reas sin necesidad de tener instrucciones explícitamente programadas. La literatura especializada considera que el término *Machine Learning* tuvo su primera aparición formal en el artículo [Samuel, 1959]. Desde su concepción, el aprendizaje automático se ha consolidado como un pilar fundamental en la investigación científica, siendo empleado en múltiples disciplinas y campos de conocimiento. Tal y como se analiza en el capítulo dedicado al estado del arte, la investigación en el campo de los (BBS) no ha permanecido ajena a dicha evolución.

Los algoritmos de *Machine Learning* se clasifican según las estrategias que emplean para aprender y procesar los datos. En la tabla 1.3 se resumen las principales categorías, acompañadas de una breve descripción y las referencias asociadas.

Tipo de Aprendizaje	Descripción
Supervisado	En este caso, el modelo es entrenado con un conjunto de datos etiquetados. Es decir, al modelo se le facilitan una serie de ejemplos, en donde se asocian una serie de entradas a sus salidas correspondientes. El objetivo es aprender una función generalizable que relacione las entradas con las salidas. [Bishop and Nasrabadi, 2006]

Continúa en la siguiente página

Tipo de Aprendizaje	Descripción
No Supervisado	Este tipo de aprendizaje busca descubrir patrones en los datos, sin necesidad de etiquetas predefinidas ni supervisión explícita. Está dirigido, por tanto, a la identificación de estructuras internas de los datos aportados mediante el descubrimiento de relaciones y agrupaciones implícitas [Hinton and Sejnowski, 1999].
Por Refuerzo	En este enfoque se basa en la toma de decisiones por parte de un agente en un entorno dinámico, a partir de recompensas y/o castigos recibidos según las acciones realizadas. El objetivo es, por tanto, desarrollar estrategias que maximicen el beneficio acumulado [Sutton, 2018].
Semisupervisado	Se trata de una de las tendencias en ML más recientes que combina datos etiquetados y no etiquetados en el proceso de entrenamiento, aprovechando la abundancia de datos sin etiquetar para mejorar la precisión del modelo cuando los datos etiquetados son escasos [Inés et al., 2024].
Autosupervisado	En este tipo de aprendizaje se generan automáticamente etiquetas a partir de los datos, permitiendo que el modelo se entrene sin necesidad de un etiquetado previo manual [Grill et al., 2020].

Tabla 1.3: Clasificación de los algoritmos de Machine Learning por categorías

Esta sección se ha centrado en ofrecer una introducción general a la inteligencia artificial y la clasificación de los algoritmos a alto nivel. No se pretende profundizar en los detalles históricos o técnicos del campo. Para explorar estos aspectos con mayor profundidad se sugiere consultar los siguientes libros: “*Artificial Intelligence: A Modern Approach*” [Russell and Norvig, 2016], “*Pattern Recognition and Machine Learning*” [Bishop and Nasrabadi, 2006], “*Machine Learning: An Algorithmic Perspective*” [Marsland, 2011] e “*Introduction to Machine Learning*” [Alpaydin, 2020].

1.4.2. Conceptos clave de los modelos deterministas frente a los probabilísticos

Una vez se han categorizado los enfoques de (*machine learning* según la naturaleza de los datos y la disponibilidad o no de etiquetas, es importante profundizar en otra perspectiva central: la separación entre métodos deterministas y probabilísticos. La clasificación en enfoques deterministas y probabilísticos refleja diferencias esenciales en cómo los algoritmos procesan la incertidumbre, estructuran el conocimiento y toman decisiones.

Característica analizada	Enfoque Determinista	Enfoque Probabilístico
Tratamiento de la incertidumbre	Directamente este enfoque no modela la incertidumbre. Por tanto, dado un conjunto de entradas específicas, siempre producen una misma salida.	Modelan explícitamente la incertidumbre mediante distribuciones de probabilidad.
Relación entre datos y salida	Fija y predecible para una entrada específica. Los resultados son reproducibles.	Se pueden dar múltiples posibles salidas asociadas a probabilidades.
Representación del conocimiento	Reglas explícitas basadas en lógica o heurísticas o parámetros optimizados aprendidos mediante algoritmos de optimización.	Distribuciones flexibles que describen el comportamiento de los datos.
Toma de decisiones	Fundamentadas en cálculos directos y predecibles.	Basada en la probabilidad asociada a cada posible resultado.
Aplicaciones típicas	Contextos con baja incertidumbre o reglas claras.	Contextos con alta incertidumbre o datos con ruido.

Tabla 1.4: Comparación entre enfoques deterministas y probabilísticos en machine learning

En el Apéndice A se presenta una tabla que sintetiza la clasificación de los algoritmos de aprendizaje automático más relevantes, proporcionando un marco de referencia que permite contextualizar la posición del algoritmo empleado en esta tesis, el estimador de máxima verosimilitud (*MLE*).

Esta clasificación permite identificar los dos enfoques predominantes en el aprendizaje automático: el estadístico, fundamentado en principios probabilísticos, y el heurístico, basado en métodos no probabilísticos o centrados en la optimización. No obstante, algunos algoritmos se consideran híbridos al combinar características de ambos paradigmas. Un ejemplo claro es el de las redes neuronales, las cuales pueden analizarse desde una perspectiva probabilística, como ocurre con las redes bayesianas, o abordarse mediante modelos deterministas estándar.

En el capítulo 1.5.2 se profundiza en mayor detalle sobre la técnica de Máxima Verosimilitud (*MLE*) aplicada en el modelo probabilístico desarrollado en esta tesis. Dado que en esta sección únicamente se ha realizado una comparación a alto nivel entre los modelos deterministas y probabilísticos, con el propósito de ofrecer una contextualización general, a continuación se presentan una serie de libros especializados en aprendizaje automático probabilístico que pueden servir como referencia por si se desea ampliar el conocimiento en esta área: ‘‘Probabilistic Graphical Models: Principles and Techniques’’ [Koller, 2009], ‘‘Bayesian reasoning and machine learning’’ [Barber, 2012], ‘‘Machine learning: a probabilistic perspective’’ [Murphy, 2012], ‘‘Probabilistic machine learning: an introduction’’ [Murphy, 2022] y ‘‘Probabilistic machine learning: Advanced topics’’ [Murphy, 2023]

1.5. Preguntas de investigación y objetivos de la tesis

El propósito de este capítulo es presentar las preguntas de investigación y los objetivos que constituyen la base de esta tesis.

Las preguntas de investigación han sido seleccionadas y formuladas para responder a los desafíos encontrados al modelar la demanda en los *BSS*. Siguiendo el planteamiento de las preguntas de investigación, los objetivos de esta tesis se dividen en un objetivo general, que resume el propósito general del trabajo, y una serie de objetivos específicos planteados a propósito para abordar las respuestas a las preguntas de forma sistemática. Para facilitar la comprensión del desarrollo de esta tesis, tanto las preguntas de investigación como los objetivos se han distribuido estratégicamente a lo largo de los capítulos. Esta organización se visualiza en la Figura 1.9.

1.5.1. Objetivo general de la tesis

En el contexto de la creciente importancia de los sistemas de movilidad sostenible, esta tesis tiene como objetivo principal **desarrollar y validar un modelo de aprendizaje automático probabilístico que permita predecir y simular la demanda de la implementación de *BSS* BiciMad**. Adicionalmente se busca identificar las variables principales que influyen en la demanda de los *BSS* y ofrecer aplicaciones prácticas del modelo que generen un impacto positivo en la planificación y explotación de los servicios en BiciMad.

Por tanto, al abordar esta problemática, la tesis pretende contribuir al avance teórico/práctico en el modelado de sistemas de movilidad sostenible, además de ofrecer soluciones aplicadas que ayuden a la operativa del servicio.

1.5.2. Preguntas de investigación - *research questions RQ's*

RQ1: ¿Cómo la evolución histórica de los *BSS*, a lo largo de sus distintas generaciones, ha influido en su gestión y, en particular, en el caso de BiciMad?

La primera pregunta de investigación se centra en comprender los *BSS* con una perspectiva histórica y de evolución. Esta investigación general es necesaria para contextualizar el caso particular de BiciMad dentro de una tendencia global más amplia. Para ello, las subpreguntas examinan, por un lado, la evolución de los *BSS* a nivel global y, por otro, las especificidades del sistema implementado en Madrid, proporcionando una visión integral del tema.

- RQ1.1: ¿Cómo han evolucionado los *BSS* desde sus inicios hasta la actualidad y cuáles son las características que definen sus generaciones?
- RQ1.2: ¿Qué particularidades definen la implementación y evolución de BiciMad en Madrid?

RQ2: ¿Cómo han evolucionado los estudios en el modelado y gestión de *BSS* desde el punto de vista de las ciencias cuantitativas y la ingeniería? La segunda pregunta de investigación tiene como objetivo analizar la evolución del conocimiento en el campo de los sistemas de bicicletas compartidas *BSS*, incluyendo la identificación de las tendencias actuales, así como los desafíos no resueltos. Se ha optado por un enfoque *top-down* que va desde la revisión sistemática de la literatura que integra los estudios generales, hasta los más específicos centrados en el problema clave del rebalanceo. Además de ofrecer una revisión sistemática del estado del arte en los *BSS*, este capítulo se centra en el diseño de una metodología específica que permita una comprensión más estructurada del campo, organizando y analizando las áreas de conocimiento existentes. Esta metodología está basada en algoritmos de clasificación, herramientas de visualización y

técnicas semánticas, con la intención de realizar una ontología de los términos disponibles en la literatura y facilitar una revisión sistemática de la misma. La pregunta principal se aborda a través de dos subpreguntas, cada una enfocada en tratar un aspecto específico del problema planteado:

- RQ 2.1: ¿Cómo se han abordado los principales desafíos? Identificando las tendencias actuales tanto de investigación, desafíos pendientes y emergentes.
- RQ 2.2: ¿Se pueden cuantificar, agrupar y categorizar las áreas de conocimiento dentro del campo de BSS de forma semántica y automática?

RQ 3: ¿Qué información pueden proporcionar los datos de BiciMad sobre el comportamiento de los usuarios y los patrones de uso del servicio?

La tercera pregunta de investigación busca realizar un análisis exploratorio de los datos (*EDA*) generados por el uso del servicio de BiciMad, y compartidos por el gestor del servicio. Este análisis es fundamental para entender la dinámica de la demanda y su relación con variables como el perfil de los usuarios, los patrones horarios de consumo del servicio y características generales de los viajes como la duración y la distancia. Por medio de las subpreguntas se abordan las dos dimensiones de la pregunta principal: La primera relacionada con la identificación de patrones propiamente dicha y la segunda enfocada en la necesidad de integrar datos de fuentes adicionales para conseguir un modelo más preciso.

- RQ 3.1: ¿Cuáles son los patrones del uso del servicio de BiciMad en términos de perfil de usuarios, horarios, duración y distancia?
- RQ 3.2: ¿Qué variables clave permiten modelar la demanda? ¿Es necesario la obtención de variables adicionales?

RQ 4: ¿Cómo se puede modelar la demanda del servicio de BiciMad?

El aspecto principal de esta investigación consiste en modelar la demanda del sistema BiciMad, considerando las incertidumbres asociadas al servicio. La cuarta pregunta de investigación aborda precisamente cómo incorporar la incertidumbre a la modelización mediante el uso de machine learning probabilístico. Las subpreguntas se enfocan primeramente en estudiar el impacto de la incertidumbre en la demanda y cómo medirlo, y posteriormente en la implementación del modelo.

- RQ 4.1: ¿Cómo afecta la incertidumbre a la demanda asociada a la realización de un viaje? ¿Es posible cuantificarla?
- RQ 4.2: ¿Se puede crear un modelo de predicción de la demanda para BiciMad que sea explicativo y contemple la incertidumbre?

RQ5: ¿Qué aplicaciones prácticas puede tener el modelo probabilístico desarrollado para la gestión y mejora del servicio BiciMad, además de la predicción de la demanda de viajes?

La capacidad predictiva del modelo probabilístico no se limita exclusivamente en la estimación de la demanda. La quinta pregunta de investigación analiza de qué manera el modelo puede ser empleado para aplicaciones adicionales. Esta pregunta otorga a la tesis un enfoque práctico, como ayuda para la gestión del servicio de BiciMad.

1.5.3. Objetivos específicos

OB1.1: Realizar un análisis histórico de la aparición y evolución de los BSS, diferenciando las distintas generaciones y detallando sus características.

El primer objetivo de esta tesis consiste en realizar una contextualización de los sistemas de bicicletas compartidas *BSS*. Para ello, se introduce el concepto de “*Mobility as a Service*” (*MaaS*) como marco teórico que ha dado lugar a diversas aplicaciones según el tipo de vehículo empleado. Los *BSS* se han consolidado como una de sus soluciones más destacadas. Se analiza el origen de estos sistemas, remontándose a su primera implementación en 1965 en Ámsterdam. Posteriormente, se describen los distintos criterios empleados para clasificar las implementaciones de *BSS* a lo largo del tiempo, concluyendo con su categorización en las diferentes generaciones, detallando las principales características de cada una.

Con este objetivo se acotará tanto el concepto de *BSS* como los términos asociados básicos para comprender la gestión de estaciones, patrones de uso y avances tecnológicos asociados, siendo la base de capítulos posteriores.

OB1.2: Analizar la implementación y evolución específica de un BSS en Madrid: caso de BiciMad.

El segundo objetivo de esta tesis, una vez contextualizados los sistemas de bicicletas compartidos en general, consiste en analizar la implementación y evolución concreta del caso particular de BiciMad. Este análisis examina su puesta en marcha en 2014, en un contexto marcado por la apuesta por la movilidad sostenible en Madrid. BiciMad representó un modelo pionero en el uso de bicicletas eléctricas en los *BSS*. Se estudian además los hitos clave de su crecimiento, como la expansión de la red, la adopción de nuevas tecnologías en las bicicletas y las transformaciones en el modelo operativo.

Los cambios significativos ocurridos en los años más recientes, como las ampliaciones de cobertura, la renovación de las bicicletas y estaciones del sistema y la reestructuración del sistema tarifario, requieren de un periodo más amplio de estabilización para poder ser evaluados. En consecuencia, estos aspectos se consideran como puntos de partida para futuros estudios y líneas de investigación. Esto no implica que el modelo propuesto no sea aplicable a los años más recientes, sino que será necesario ajustar ciertos parámetros para reflejar la nueva realidad del sistema y capturar adecuadamente las dinámicas actuales.

OB2.1: Desarrollar una metodología basada en algoritmos no supervisados y técnicas de visualización para crear una ontología de los términos disponibles en la literatura y facilitar una revisión sistemática en BSS.

El desarrollo de una ontología para reducir la fragmentación de los términos disponibles en la literatura sobre sistemas de bicicletas compartidas *BSS* es un paso crucial para estructurar el conocimiento en este campo de conocimiento. Además, fue necesario implementar un enfoque que permitiera agrupar en temáticas de las investigaciones realizadas en los *BSS* mediante técnicas de análisis no supervisado y herramientas de visualización. La metodología desarrollada permitió, por un lado, identificar las tendencias de investigación en los diversos periodos temporales y por otro facilitar la posterior revisión sistemática del estado del arte.

OB2.2: Realizar una revisión sistemática del estado del arte en los BSS, con un enfoque que abarca desde los estudios generales en BSS hasta aquellos que se centran en uno de sus principales problemas: el rebalanceo.

Este objetivo de la tesis se centra en la realización de una revisión sistemática del estado del arte en sis-

temas de bicicletas compartidas. Esta revisión tiene como propósito sintetizar y organizar la literatura existente, ofreciendo una visión integral que abarca desde los estudios generales sobre los BSS como aquellos que analizan específicamente las técnicas y herramientas utilizadas para abordar el problema del rebalanceo. Esta problemática, propia de la mayoría de los sistemas de bicicletas compartidas, trata sobre la necesidad de mantener un equilibrio dinámico entre la oferta y la demanda de bicicletas en las estaciones.

Con este objetivo se consigue contextualizar la investigación realizada en esta tesis dentro del conocimiento existente en este campo y justificar su relevancia. Se identifican lagunas en el conocimiento, se delimita el alcance de la investigación y se localizan en la literatura enfoques probados y comparables, permitiendo destacar la contribución original de la tesis.

OB3.1: Realizar un análisis exploratorio y descriptivo de los datos de viajes del servicio BiciMad. Identificando patrones de uso y determinando variables que caracterizan la demanda.

El análisis exploratorio y descriptivo de los datos de viajes del servicio BiciMad permite comprender los patrones de uso y las dinámicas de demanda en el servicio. El objetivo consiste en analizar los datos disponibles, identificando tendencias clave como los horarios de mayor afluencia, las estaciones más demandadas y las características de los trayectos realizados. Más allá de describir los comportamientos de los usuarios, el objetivo subyacente es el de identificar cómo condicionan la demanda, factores externos tales como el clima, el horario o la disposición geográfica de las estaciones.

OB3.2: Combinar variables existentes con otras de diferentes fuentes de información para el modelado de la demanda de viajes en el servicio de Bicimad.

Este objetivo busca enriquecer el dataset inicial de viajes de BiciMad con el fin de mejorar la capacidad de modelar la demanda del sistema. El enfoque seguido para alcanzar este objetivo se basa en dos líneas de acción. La primera consiste en combinar las variables existentes para la obtención de variables adicionales, como puede ser el caso de obtener la distancia recorrida a partir de los valores de latitud y longitud de las estaciones origen y destino. La segunda permite integrar datos operativos del sistema con variables externas como son las condiciones meteorológicas o el calendario de festivos. Este enriquecimiento de los datos originales permitirá que la demanda se modele de manera más precisa, al capturar factores contextuales que influyen en el uso del sistema.

OB4.1: Desarrollar un modelo de ML probabilístico que prediga la demanda del servicio de bicicletas BiciMad a partir de los datos empíricos de viajes de 2018.

El propósito de este objetivo consiste en capturar tanto los patrones observados como la incertidumbre inherente a los datos en BiciMad, mediante el desarrollo de un modelo de aprendizaje automático de naturaleza probabilística. Los modelos probabilísticos ofrecen una ventaja significativa al permitir no solo estimar valores puntuales de demanda, sino también generar distribuciones que reflejen la variabilidad y posibles escenarios futuros o contrafactuales, lo que incrementa la capacidad de tomar decisiones informadas, orientadas a mejorar el servicio de BiciMad. La construcción de este modelo se fundamentará en los datos enriquecidos previamente.

OB4.2: Validar el modelo de machine learning probabilístico frente a datos empíricos del servicio de BiciMad de 2019

El propósito de este objetivo es evaluar la capacidad predictiva y la robustez del modelo de Machine Learning probabilístico desarrollado previamente, utilizando los datos empíricos del servicio BiciMad correspondientes al año 2019. Este proceso de validación garantizará que el modelo no solo se ajusta a los datos históricos de 2018, sino que es capaz de generalizar sus predicciones, permitiendo comparar las predicciones

con los datos empíricos correspondientes al año 2019.

Como se ha mencionado previamente, la validación queda restringida al año 2019, al haberse dado una serie de eventos en los años posteriores que influyeron significativamente en la demanda del servicio y el comportamiento general de los usuarios.

OB5.1: Demostrar la capacidad del modelo para la imputación de datos y la cuantificación de costes ante suspensiones del servicio.

El propósito de este objetivo es el de describir y demostrar dos aplicaciones prácticas y a su vez complementarias del modelo de Machine Learning desarrollado, orientadas a optimizar el servicio de BiciMad: la imputación de datos faltantes y la estimación de los costes derivados de interrupciones en el servicio.

La imputación de datos es esencial en la analítica de datos para garantizar la calidad de los análisis y predicciones. Si bien existen técnicas genéricas descritas en la literatura para la imputación de datos faltantes, en este caso y mediante el análisis de patrones históricos y relaciones entre variables operativas que realiza el modelo, se pueden reconstruir datos ausentes que sean coherentes con las distribuciones probabilísticas esperadas de los datos.

Adicionalmente, el modelo permite cuantificar el impacto económico de interrupciones en el servicio basándose en las expectativas de demanda en condiciones normales sin interrupciones.

OB5.2: Cuantificar la demanda no cubierta y no registrada por el sistema actual.

Uno de los retos actuales que presenta el servicio de BiciMad es el de la imposibilidad de identificar y cuantificar la demanda no cubierta y no registrada por el sistema. La demanda no cubierta incluye situaciones en las que el usuario no puede iniciar un viaje por falta de bicicletas disponibles en la estación origen, así como aquellas en las que no puede finalizar su trayecto al no haber esta vez anclajes libres para dejar la bicicleta en la estación destino.

En este objetivo se sientan las bases de una metodología que permite cuantificar la demanda no cubierta, comparando escenarios alternativos generados por el modelo con la disponibilidad de bicicletas y anclajes en las distintas estaciones en tiempo real.

Los resultados de esta aplicación práctica permitirían identificar áreas críticas del sistema y formular recomendaciones para mejorar la distribución de bicicletas, la capacidad de las estaciones y la planificación estratégica del servicio.

OB5.3: Verificar la capacidad predictiva del modelo mediante la detección de datos erróneos o anomalías en los datos empíricos con frente a lo esperado por el modelo.

El propósito de este objetivo es detectar datos del conjunto de datos empíricos del servicio de BiciMad, que no se ajusten con los patrones esperados generados por el modelo. Las anomalías pueden surgir por múltiples razones, como fallos en la captura de datos, errores humanos o eventos externos que desvían los patrones de uso habituales. Se aplicará el modelo probabilístico para comparar las distribuciones de datos observadas con las predicciones esperadas, empleando técnicas de análisis de discrepancias.

Como se explicará en un capítulo posterior, este objetivo permitió detectar información errónea en las horas asignadas a ciertos viajes del año 2019. Este hecho fue validado posteriormente por el administrador de datos del servicio, demostrando la capacidad del modelo para detectar posibles errores en los datos y garantizar la calidad de los mismos.

OB5.4 Desarrollar un simulador para evaluar escenarios contrafactuales, resultantes de la adición o eliminación de estaciones y optimizar la toma de decisiones.

Esta herramienta de simulación propuesta permitirá evaluar el impacto de las modificaciones efectuadas

en la red de estaciones. Al ser capaces de modelar escenarios hipotéticos a partir de datos históricos y predicciones basadas en el modelo probabilístico, el simulador permitirá a los gestores del servicio anticipar las consecuencias de decisiones estratégicas y mitigar posibles riesgos operativos.

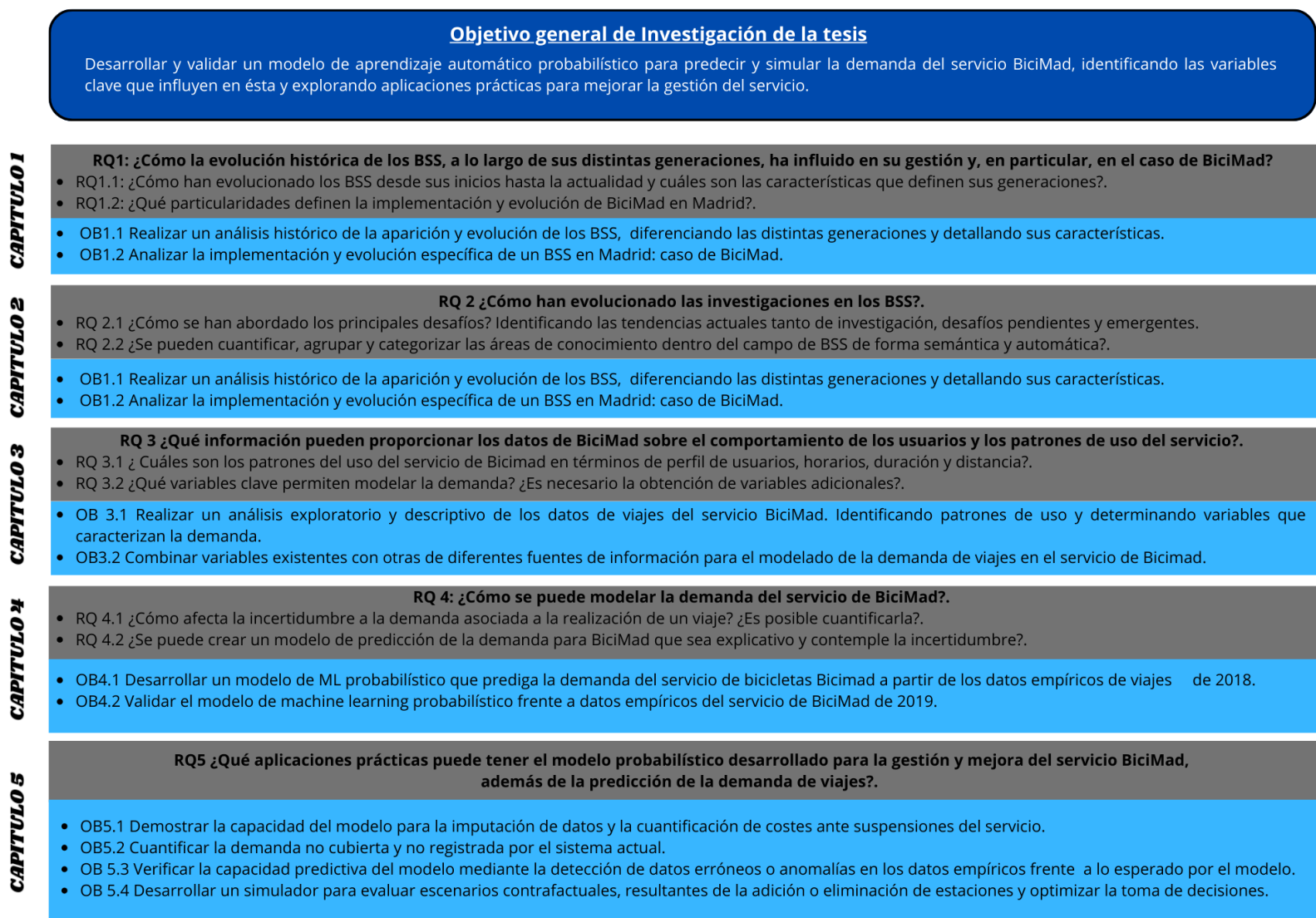


Figura 1.9: Representación esquemática de la relación entre las preguntas de investigación, el objetivo principal, los objetivos secundarios y la estructura en capítulos de la tesis. Fuente: Elaboración propia.

1.6. Esquema de la tesis doctoral

El contenido de esta tesis se desarrolla a lo largo de seis capítulos, distribuidos de la manera que se detalla a continuación:

- **Capítulo 1 Introducción y Motivación:**

El primer capítulo establece el contexto general del ámbito de estudio de esta tesis: los sistemas de bicicletas compartidas *BSS*. Se inicia con una introducción a los conceptos clave de la movilidad urbana, seguida de un análisis histórico de la evolución de los *BSS* a lo largo del tiempo. Dado que la investigación se centra específicamente en la implementación del servicio de BiciMad, se realiza a continuación una contextualización de dicha implementación. Posteriormente se presenta la motivación de la investigación junto con el objetivo principal y los objetivos secundarios, los cuales se relacionan directamente con las preguntas de investigación. Finalmente se proporciona una descripción de la estructura del documento para facilitar su lectura y comprensión.

- **Capítulo 2 Estado del Arte:**

En el segundo capítulo se muestra una revisión detallada del estado del arte con el fin de definir un marco teórico y conceptual en el cual basar la tesis. Para ello, se elabora una metodología que facilita la identificación y el análisis de estudios relevantes para el campo de estudio, abordando en dicho análisis la fragmentación terminológica en la literatura mediante técnicas de agrupación por temas, la creación de un tesoro fundamentado en una ontología terminológica y el uso de herramientas de visualización.

A continuación, se analizan los artículos de revisión clave del campo de estudio, así como los principales hitos e investigaciones relacionadas con los sistemas de bicicletas compartidas (*BSS*). Este análisis tiene como objetivo identificar tanto las contribuciones más significativas como los desafíos pendientes en el ámbito. La revisión sigue un enfoque *top-down*), comenzando por investigaciones generales sobre los *BSS* y culminando en el análisis detallado del problema principal abordado en esta tesis: el rebalanceo de bicicletas.

- **Capítulo 3 Análisis exploratorio de los datos (EDA):**

Este capítulo comienza describiendo las fuentes de datos utilizadas y los datasets obtenidos, proporcionando una explicación detallada de los campos incluidos. A continuación, se detallan los procesos de extracción y transformación realizados para preparar los datos a partir de las fuentes originales, basándose en un análisis exploratorio de datos (*EDA*). Dicho análisis tuvo como objetivo comprender, explorar y preparar el conjunto de datos, identificando patrones, relaciones entre variables, valores atípicos y datos faltantes, así como evaluando características clave y distribuciones iniciales. Finalmente, se presenta el dataset resultante que será utilizado en la construcción del modelo propuesto.

- **Capítulo 4 Enfoque de ML Probabilístico para el Modelado de la Demanda en BiciMad** El capítulo cuarto se centra en describir e implementar un modelo de *machine learning* probabilístico con el fin de representar y predecir la demanda del servicio de BiciMad. Para ello, se centra en el análisis de la distancia recorrida y el número de viajes por hora, ajustándolos a distribuciones probabilísticas. Adicionalmente en el capítulo se detalla la creación de un simulador basado en dicho modelo, el cual permite generar escenarios hipotéticos pero siempre alineados con las distribuciones ajustadas a los datos empíricos.

- **Capítulo 5 Aplicaciones del Simulador - Modelo**

A lo largo de este capítulo se enumeran, describen y se explica la implementación de posibles aplicaciones prácticas del modelo probabilístico desarrollado en el Capítulo 4. El objetivo es demostrar que la utilidad de un modelo probabilístico trasciende el simple hecho de explicar en términos de probabilidades un modelo complejo. Para ello, se detallan en este capítulo utilidades prácticas de la capacidad que tiene el modelo de generar datos sintéticos, las cuales pueden ayudar a BiciMad a la gestión y optimización del servicio.

■ **Capítulo 6 Conclusiones, contribuciones y futuros desarrollos**

Finalmente, en este capítulo se resumen las principales conclusiones obtenidas a partir del desarrollo de esta tesis. También se describe para cada una de ellas su importancia y su aporte al estado del arte. Adicionalmente se exponen los artículos publicados o en proceso de publicación que forman parte de la tesis. Seguidamente se mencionan otros méritos, como artículos publicados pero fuera del marco de la tesis y participaciones en conferencias. Finalmente se plantean posibles líneas de investigación futuras para continuar y profundizar en las investigaciones derivadas de esta tesis.

“Daher ist die Aufgabe nicht sowohl zu sehen was noch keiner gesehen hat, als bei Dem was Jeder sieht, zu denken was noch Keiner gedacht hat”.

Por tanto, la tarea no es tanto ver lo que nadie ha visto antes, sino pensar en lo que nadie ha pensado todavía en lo que todos ven.

— Arthur Schopenhauer *Parerga y Paralipomena. Escritos filosóficos menores (Parerga und Paralipomena. Kleine philosophische Schriften)*, 1851

2

Estado del arte

Índice del capítulo

2.1. Revisión del estado del arte y descripción de la metodología	34
2.2. Identificación de palabras clave	35
2.3. Inventario de artículos de revisión	37
2.4. Fases y áreas clave en la evolución de la investigación en BSS	40
2.5. El problema del rebalanceo	45
2.6. Conclusiones del capítulo	52

Este capítulo se centra en presentar un análisis detallado del “Estado del arte” en los sistemas de bicicletas compartidas (**BSS**). Se revisan los hitos más relevantes en la evolución del conocimiento en este campo, permitiendo identificar las principales contribuciones y lagunas. Por tanto, a través de esta revisión se establece el marco de conceptual y teórico que supone la base de esta tesis, permitiendo el desarrollo del resto de capítulos. Las preguntas a las cuales da respuesta este capítulo son: *¿Qué investigaciones se han realizado en el campo de conocimiento de los BSS? ¿Qué limitaciones y resultados tienen? ¿Qué lagunas de investigación se identifican? ¿Dónde se sitúan las aportaciones de la tesis dentro de la literatura?*

2.1. Revisión del estado del arte y descripción de la metodología

El campo del BSS engloba distintos problemas (operacionales, económicos, ecológicos, ...) por lo que la literatura científica a este respecto es demasiado amplia para el ámbito de una tesis en ingeniería. En este contexto, para poder sistematizar el “Estado del Arte” se implementó una metodología sistemática en la recopilación, selección y análisis de las referencias bibliográficas. Esta forma de abordar la revisión de la literatura es común en otras áreas de la ciencia (típicamente aquellas relativas a las Ciencias de la Salud) pero no tanto en Ingeniería. Las revisiones sistemáticas ofrecen numerosos beneficios: Proporcionan una visión objetiva de la evidencia disponible sobre un tema específico, minimizando los sesgos a través de métodos rigurosos.

Al reunir información de múltiples estudios independientes, estas revisiones aumentan la fiabilidad y generalizabilidad de los hallazgos, identifican brechas en la literatura y guían a los investigadores hacia áreas que requieren más estudio. La evaluación de la calidad y la transparencia en el proceso de revisión sistemática aseguran la precisión y replicabilidad de los resultados, haciendo de estas revisiones herramientas esenciales para el avance del conocimiento, por lo que este tipo de trabajos tiene valor científico *intrínseco*. Por estas razones, hemos abordado este capítulo desde esta perspectiva metodológica.

En primer lugar, se identificaron las palabras clave más adecuadas, que estuvieran alineadas con el objeto de estudio de esta investigación. A continuación se realizaron búsquedas exhaustivas en bases de datos académicas reconocidas (*Web of Science*, *Scopus* y *Google Scholar*), aplicando filtros específicos (e.g., fechas, calidad de la revista, número de citas) para refinar los resultados. Entre de los resultados obtenidos se priorizó la identificación de “*Review Papers*” los cuales ofrecen una visión general del estado del conocimiento y un resumen de las investigaciones previas en el área. Seguidamente, se seleccionaron los artículos más significativos, que fueron sometidos a una lectura y un análisis detallado, destacando aquellos con mayor impacto académico (en términos de citas y contribuciones). Los artículos seleccionados fueron organizados de manera temática y cronológica para tener una visión clara y estructurada de la evolución del conocimiento. Además, se aplicó la técnica “*snowballing*” para ampliar la revisión mediante el análisis de las referencias citadas en los artículos clave, con el fin de identificar estudios adicionales que no fueron detectados en una primera búsqueda y filtrado.

2.2. Identificación de palabras clave

Para seleccionar las palabras clave, que nos permitan posteriormente realizar búsquedas en bases de datos académicas relevantes, hemos optado por hacerlo con un enfoque *top-down*.

Como primera palabra clave se ha seleccionado el término *bike-sharing* por dos razones principales: (1) por tratarse del concepto más amplio dentro del marco de esta investigación (permitiendo abordar tanto aspectos teóricos como aplicaciones prácticas) y (2) por facilitar una visión global del estado del conocimiento, a la vez que contextualiza nuestra investigación, al integrar enfoques tecnológicos, económicos, sociales, de optimización y de sostenibilidad.

Siguiendo el enfoque adoptado para la identificación de palabras clave, resulta esencial seleccionar términos más específicos alineados directamente con los objetivos de investigación descritos en esta tesis. Se ha decidido incluir el acrónimo *BSS* (“*Bike-Sharing Systems*”), por abarcar el conjunto de sistemas de bicicletas compartidas. Asimismo, se seleccionan “*dock-based*” como una subcategoría clave dentro de los *BSS* y “*rebalancing problem*” como uno de los principales desafíos abordados en el campo de estudio de estos sistemas. En esta tesis se aborda un subproblema específico del rebalanceo: El modelado de la demanda. Un aspecto crucial para optimizar la resolución del problema del rebalanceo, y por extensión para mejorar la operativa de los *BSS*.

Al realizar las primeras búsquedas en bases de datos académicas con los términos seleccionados, se identificó una problemática frecuente en la literatura relacionada con los *BSS*: la falta de homogeneidad en la terminología empleada. Se observó que diferentes autores no solo emplean sinónimos, sino también nomenclaturas alternativas o variaciones en la escritura (como “*bike sharing*”, “*bikesharing*” o “*bike-sharing*”) para referirse a los mismos conceptos clave. Esta disparidad y falta de consistencia terminológica dificulta la localización de artículos relevantes, fragmentando el acceso al conocimiento. En la Tabla 2.1 mostramos un ejemplo de cada tipo de discrepancia encontrada.

Concepto clave	Variantes encontradas	Problema identificado
Bike sharing	Bike-sharing, bikesharing, bike sharing, bike-sharing	Variaciones en la forma de escritura que dificultan la recuperación precisa de artículos relevantes.
Rebalancing	Redistribution, Fleet management, Relocation, Inventory distribution, Repositioning, Capacity management, Balancing, Imbalancing.	Uso de sinónimos que se solapan, generando confusión sobre si se refieren al mismo concepto.
Bike Sharing Systems (BSS)	Public Bicycle Sharing Schemes (PBSS), Bicycle Rental Systems, Urban Bike-sharing Programs, Shared Mobility Bicycle Services, Bicycle-sharing Networks, Bike Rental Services	Falta de estandarización en el uso de términos relevantes, lo que fragmenta la literatura existente.

Tabla 2.1: Ejemplos de falta de uniformidad terminológica en la literatura sobre BSS

La existencia de esta problemática reveló una brecha en la literatura existente, hecho que se tradujo en una oportunidad para realizar una contribución significativa. Como respuesta a esta necesidad desarrollamos un tesoro¹, el cual constituye la primera contribución de esta tesis. Este tesoro tiene como objetivo estandarizar la terminología en inglés utilizada en los BSS, facilitando la búsqueda y recuperación de artículos relevantes. La Figura 2.1 expone una sección del tesoro elaborado, resaltando las relaciones conceptuales entre varios diversos términos y el concepto principal de BSS. La primera versión de este tesoro fue publicada en el artículo inicial [Vallez et al., 2021] que forma parte de esta tesis doctoral. Como se detalla en dicho artículo, el tesoro también ha sido puesto a disposición de la comunidad académica y del público general, mediante un [repositorio de GitHub](#) donde se puede consultar y descargar libremente.

¹Un tesoro es una herramienta que permite organizar y estandarizar los términos relacionados con un campo de estudio, proporcionando sinónimos, variantes gráficas y jerarquías conceptuales. Entre sus principales ventajas radica la de reducir la confusión causada por la inconsistencia terminológica, que es precisamente el problema encontrado.

BSS-rebalancing-review / ThesaurusBSSKeywords.txt		
Code	Blame	732 lines (732 loc) · 12.9 KB
76	bike share program	BSS
77	bike share system	BSS
78	bike sharing	BSS
79	bike sharing operator	BSS
80	bike sharing problem	BSS
81	bike sharing program	BSS
82	bike sharing scheme	BSS
83	bike sharing service	BSS

Figura 2.1: Sección del Tesauro. Fuente: Elaboración propia.

2.3. Inventario de artículos de revisión

Para llevar a cabo un análisis riguroso del estado del arte, en el campo de los *BSS* realizamos una búsqueda de los artículos más relevantes categorizados como “*review papers*”, empleando las palabras clave previamente identificadas. Esta práctica, común en la investigación académica, nos proporciona una visión global del conocimiento acumulado, permitiéndonos identificar las principales líneas de investigación y los elementos pendientes por cubrir en la literatura existente.

Como parte de los trabajos de revisión encontrados, el artículo [DeMaio, 2009] se considera como el punto de partida de la investigación sobre los sistemas *BSS*, aportando un análisis histórico desde los primeros sistemas en los años 60 hasta la implementación de los sistemas de tercera generación. Este artículo también aborda temas como el impacto del uso de la bicicleta, la conectividad con otros medios de transporte y los diferentes modelos de provisión del servicio. Sin embargo, desde el punto de vista del interés de esta tesis, su contribución más destacable es la identificación de la distribución de las bicicletas entre estaciones (para equilibrar la oferta y la demanda) como el principal desafío operacional del servicio. Seguidamente, [Fishman et al., 2013] en su síntesis de la literatura hasta ese momento, destaca la proliferación de nuevos sistemas en Europa y Asia principalmente, y apunta que la mayoría de las investigaciones del momento estaban enfocados a aspectos relacionados con la operación y el uso del servicio. También incide en la necesidad orientar investigaciones futuras a temas relacionados con la sostenibilidad y la seguridad. Posteriormente, en el artículo [Fishman, 2015], se revisan las nuevas investigaciones incorporadas hasta 2015 poniendo el foco en dos aspectos principales. (1) el crecimiento de los servicios a nivel global y (2) la incorporación de tecnologías avanzadas como bicicletas eléctricas o sistemas sin necesidad de anclaje. Siguiendo con los artículos de revisión en [Ricci, 2015] se recopilan de artículos relacionados con el impacto de los *BSS*, con especial interés en aspectos clave como la salud pública y el medioambiente. Además, se incluyen otros tipos de estudios centrados en la implementación y operación de estos sistemas, lo cual resulta básico y relevante para el estudio integral de los *BSS*. Con relación a los objetivos de esta tesis, destacan las aportaciones relacionadas con el reequilibrio de la demanda y oferta de bicicletas en las diversas estaciones. Entre esas aportaciones son de especial interés: (1) El reconocimiento del costo y complejidad del proceso de reequilibrio, (2) La relación

entre la ubicación y la demanda, subrayando que la colocación estratégica de estaciones reduce la necesidad de reequilibrio y (3) La necesidad de innovación en los métodos de reequilibrio, sugiriendo el empleo de vehículos no contaminantes para reducir el impacto negativo en las emisiones. Por otra parte, y siguiendo la línea cronológica, está el artículo [Si et al., 2019]. Se trata de una revisión exhaustiva de la literatura entre los años 2010 y 2018, integrando diversas áreas de investigación del campo de los BSS y con un enfoque cientimétrico². En este artículo se hace especial énfasis en los modelos emergentes *dockless* de BSS, si bien hay que aclarar que este subtipo de sistemas de bicicletas compartidas queda fuera del alcance de esta tesis.

Aunque revisiones anteriores como las de [DeMaio, 2009; Fishman et al., 2013; Fishman, 2015; Ricci, 2015] han contribuido significativamente a la estructuración de la investigación y el conocimiento sobre los BSS, al iniciar el análisis del estado del arte se identificaron brechas importantes en la literatura, que o bien habían quedado fuera del alcance o no se habían tratado con suficiente detalle en los artículos de revisión existentes. Estas brechas se pueden agrupar en tres áreas principales:

1. Temporal, ya que no se habían incluido investigaciones publicadas entre 2018 y 2020.
2. Terminológica, con una fragmentación de términos que generaba confusión (tal y como se recoge en [Vallez et al., 2021] y motivó la creación por nuestra parte de un tesoro específico.
3. Metodológica, remarcando la falta de un análisis más exhaustivo de las soluciones aportadas en el problema del rebalanceo. Todo ello puso de manifiesto la necesidad de un análisis metodológico más profundo y detallado de la literatura existente. Esta necesidad fue cubierta en el artículo de nuestra autoría [Vallez et al., 2021], el cual forma parte de esta tesis.

Es importante destacar en este punto que la ampliación de la revisión de la literatura se lleva hasta 2020, por ser éste el año en el cual desarrollamos dicho artículo, siendo finalmente publicado en 2021. Las principales aportaciones de nuestro artículo [Vallez et al., 2021] en el campo de los BSS pueden resumirse en los siguientes puntos clave:

- Tesoro específico para los sistemas de bicicletas compartidas: Desarrollamos un tesoro que ayuda a unificar la nomenclatura. Este recurso, facilita la identificación y la búsqueda de estudios relevantes, y ayuda a resolver los problemas de variabilidad terminológica existente en este campo de estudio.
- Clusterización en temáticas de las investigaciones realizadas en los BSS: Hemos agrupado y organizado la literatura científica sobre BSS. Clasificamos los diversos estudios disponibles en el área de conocimiento de los BSS mediante técnicas de análisis no supervisado y herramientas de visualización, en concreto VOSViewer [Van Eck and Waltman, 2010]. Esto nos permitió identificar tendencias de investigación variantes a lo largo de diversos períodos temporales.
- Enfoque en el problema del “rebalancing” de bicicletas en los BSS: Realizamos una clasificación exhaustiva de las técnicas de rebalanceo, dividiéndolas entre estáticas, basadas en horarios predefinidos, y dinámicas que, al contrario que las anteriores, se reajustan en función de la demanda real. Además, organizamos los artículos en función de los algoritmos empleados para redistribuir las bicicletas. Una sección posterior de esta revisión 2.5 profundiza en el problema del rebalanceo, aportando una revisión más exhaustiva de la literatura relacionada, presenta una síntesis comparativa de los trabajos en forma de tabla resumen con las diversas técnicas empleadas y la identificación de una carencia que constituye una aportación original de esta tesis.

²La traducción más aceptada de “*scientometric*” al castellano es *cientimétrico*. La *cienciometría* es la disciplina que emplea indicadores cuantitativos para medir y evaluar el rendimiento de la investigación científica. Aporta un análisis sobre el desarrollo y el impacto en general de la ciencia

Después de la publicación de nuestro artículo, el campo de los *BSS* ha continuado su evolución, dando lugar a la aparición de nuevos artículos de revisión. En el artículo [Zhou et al., 2022] los autores realizaron una revisión de las publicaciones entre 2010 y 2020, nuevamente con un enfoque centimétrico combinado con un análisis bibliométrico³. El análisis bibliométrico facilitó la identificación de patrones de citación y colaboración entre autores, mientras que el centimétrico aportó un análisis de las tendencias y áreas emergentes en el campo de conocimiento de los *BSS*. Es destacable la aportación que realiza esta revisión en dos áreas principales: (1) la sostenibilidad y (2) el uso del aprendizaje automático en la predicción de la demanda del servicio de los *BSS*. Hay que recordar que la predicción de la demanda es un subproblema clave en la resolución del desafío del rebalanceo en los *BSS*.

Posteriormente, se han desarrollado revisiones centradas en el problema del rebalanceo y las diferentes estrategias empleadas para la recolocación de los vehículos. El artículo [Zhang et al., 2023] realiza una revisión integral de la reubicación en los sistemas de vehículos compartidos (*“Vehicle-Sharing Systems”* (*VSS*)), concretamente en los *“Car-Sharing Systems”* (*CSS*), *“Bike-Sharing Systems”* (*BSS*) y los *“Scooter-Sharing Systems”*. Los autores realizan una clasificación jerárquica de las estrategias de reubicación que abarcan desde la planificación del sistema hasta la explotación diaria. Comentan también que las estrategias pueden estar basadas en los usuarios o en la gestión de los operadores y propone que futuras investigaciones se orienten a una mezcla de ambos enfoques. Al tratar distintos tipos de *VSS* permite establecer futuras sinergias entre diferentes sistemas a la hora de abordar el problema del rebalanceo. Finalmente, destacar de este artículo la introducción de una discusión novedosa: los problemas específicos de los *“Electric Car-Sharing Systems”* (*ECSS*) relaciones con la optimización de la carga. Estos mismos problemas pero a otra escala afectan a las últimas generaciones de *BSS* las cuales han incorporado bicicletas eléctricas a su flota. Complementando al anterior artículo está [Teusch et al., 2023] en el cual los autores se centran en una revisión exhaustiva de los artículos, comprendidos entre 2012 y 2022, que describen el impacto del aprendizaje automático en los sistemas de movilidad compartida *“Shared Mobility Systems”* (*SMS*). A diferencia de los autores anteriores, en esta recopilación se concentran exclusivamente técnicas de machine learning pero aplicadas tanto a la parte del modelado de la demanda como a la optimización de la reubicación de los vehículos, ambos subproblemas del ampliamente reconocido problema del rebalanceo de vehículos en estos sistemas.

A pesar de que el artículo que se analiza a continuación tiene una fecha de publicación anterior, se ha optado por introducirlo después para seguir un enfoque estructurado que prioriza el análisis desde lo más amplio, los *SMS*, hasta lo más específico, los *BSS*.

Para concluir con el análisis del inventario de artículos de revisión relevantes para esta investigación, es necesario referenciar a [Albuquerque et al., 2021]. En este artículo, los autores realizan una revisión sistemática de las investigaciones relacionadas con la aplicación a los *BSS* de técnicas de aprendizaje automático, publicadas entre 2015 y 2019. En su análisis destacan técnicas tanto de clasificación como de predicción, tales como: Redes Neuronales Recurrentes *“Recurrent Neural Networks”* (*RNNs*) incluyendo las redes de memoria a corto-largo plazo *“Long Short -Term Memory networks”* (*LSTM*), *“Radom Forest”*, Clustering Jerárquico y *“K-Means”*, Redes Neuronales Convolucionales *“Convolutional Neural Networks”* (*CNN*) y Algoritmos de Montecarlo.

Esta sección ha ofrecido una visión global, siguiendo la línea temporal, de los artículos de revisión *“review papers”* incluyendo tanto aquellos centrados en *BSS* desde una perspectiva general como aquellos más centrados en la aplicación de técnicas de *“machine learning”* para aportar soluciones al problema del rebalanceo. La Tabla D.1 del apéndice D resume los artículos mencionados con sus principales características. En los próximos subcapítulos se ofrece un análisis detallado de los artículos del campo de conocimiento de los

³La bibliometría (en inglés *“bibliometric”*) es considerada una subdisciplina dentro de la centimetría (*“scientometric”*) que se encarga del análisis cuantitativo de las publicaciones científicas. Emplea índices como número de citas, la citación conjunta de documentos o la ocurrencia de palabras clave, con el fin de evaluar la producción científica y la conexión entre investigaciones.

BSS que no son revisiones.

2.4. Fases y áreas clave en la evolución de la investigación en BSS

La notable proliferación de publicaciones científicas en el ámbito de los sistemas de bicicletas compartidas (*BSS*) se origina principalmente a partir del año 2010 [Si et al., 2019; Vallez et al., 2021].

Al tratarse de un campo de investigación relativamente moderno y que a su vez ha recibido un alto interés por la comunidad científica, se ha producido una diversificación en los enfoques y temáticas abordadas a lo largo de los años.

En el caso de *BSS*, de acuerdo con lo señalado por [Si et al., 2019], las tendencias temáticas dentro de la investigación en *BSS* pueden clasificarse en cuatro etapas temporales claramente definidas:

- 2010 - 2012: En este período la investigación se enfocó principalmente en temas relacionados con la seguridad vial y el desarrollo de políticas públicas, con el objetivo de garantizar la implementación efectiva de los *BSS* y promover su integración en el sistema de transporte urbano.
- 2013 - 2014: En estos años la temática de investigación en los *BSS* se enfocó en tres ideas clave: los beneficios asociados al uso de estos sistemas, su impacto en diferentes dimensiones como la movilidad urbana, el medio ambiente y la salud pública y finalmente en el diseño y funcionamiento de los propios sistemas.
- 2015: Comienza la focalización de la investigación en aspectos fundamentales del servicio como son la optimización de los sistemas, el análisis del comportamiento de los usuarios, la interacción con el entorno construido, el diseño eficiente de las implementaciones de *BSS* comparando casos de uso de éxito con aquellos que presentaban ciertos problemas y el desarrollo de infraestructura adecuada para su implementación y sostenibilidad
- 2016 -2018: Los estudios se orientaron en este período principalmente a áreas como el análisis de la demanda, la optimización del rebalanceo y la redistribución de bicicletas, las preferencias, hábitos de uso y percepción del servicio de los usuarios, la influencia de variables externas como el clima y la interacción con otros medios de transporte. Se identifica además un área de investigación destacada relacionada con los sistemas de bicicletas compartidas (*BSS*) sin estaciones fijas (dockless). En nuestro artículo [Vallez et al., 2021], presentamos la Figura 2.2, que muestra de manera clara cómo la investigación en los sistemas de bicicletas compartidas ha evolucionado con el tiempo, alcanzando en 2019 un enfoque predominante en los sistemas dockless.
- 2019 -2024: En los últimos años, la investigación ha retomado el enfoque en los sistemas de bicicletas compartidas basados en estaciones (dock-based), desarrollando soluciones innovadoras para abordar los desafíos del rebalanceo y la recolocación de bicicletas. Estas soluciones se han potenciado mediante el uso de herramientas avanzadas basadas en inteligencia artificial, con un énfasis particular en enfoques de reposicionamiento dinámico. Además, se observa una tendencia creciente hacia la creación de modelos capaces de simular las condiciones operativas del servicio, permitiendo un análisis más profundo y una planificación más eficiente. La Figura 2.3 muestra una actualización de la imagen publicada previamente en nuestro trabajo en la Ref. Vallez et al. [2021].

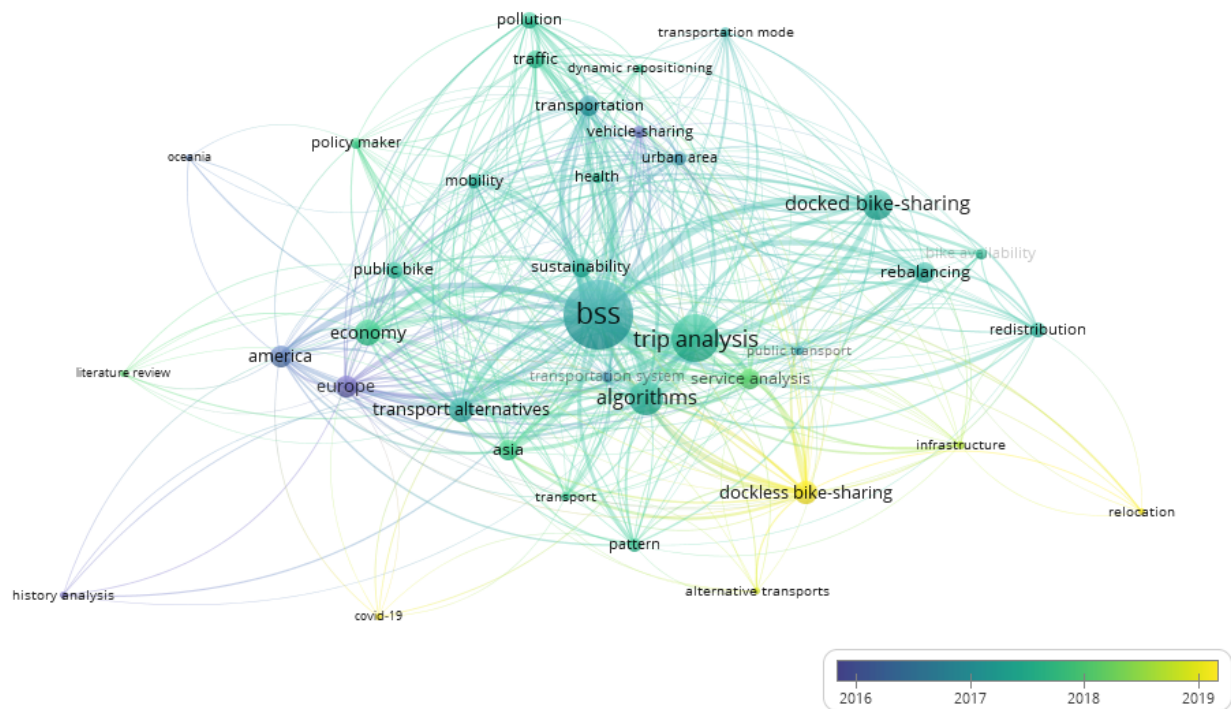


Figura 2.2: Evolución temporal de la investigación en BSS por temáticas hasta 2019. Fuente: Elaboración propia publicada en [Vallez et al., 2021]

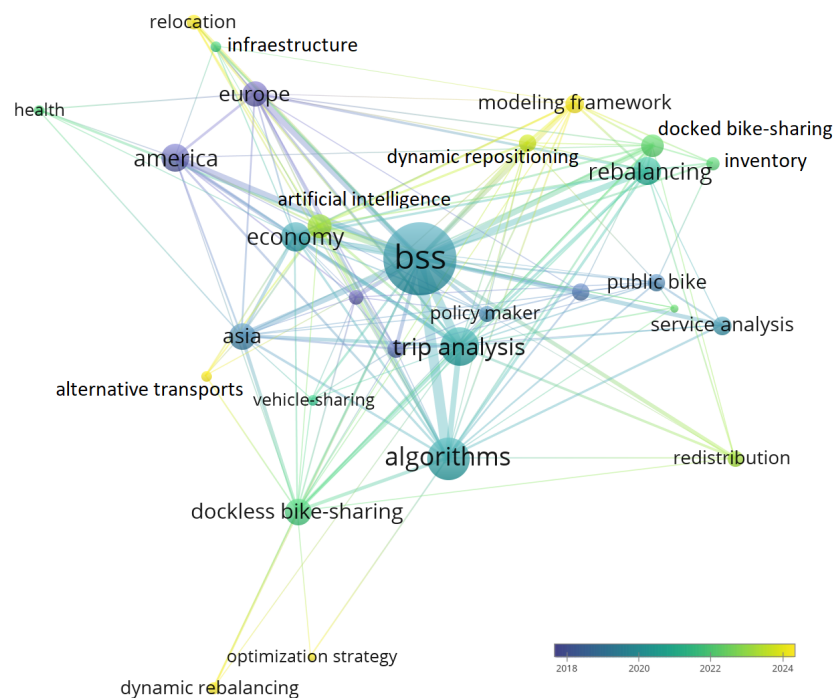


Figura 2.3: Evolución temporal de la investigación en BSS por temáticas hasta 2019. Fuente: Elaboración propia publicada en Vallez et al. [2021]

Una vez mostrada la evolución temporal de la investigación en BSS, el siguiente paso consiste en agrupar los temas de investigación. Para ello, se parte de la representación gráfica de las principales palabras clave y temas que aparecen en los títulos y resúmenes de los artículos relacionados con el bikesharing mostrada en la Figura 2.3. Utilizando el mismo software libre (VOSviewer) empleado para la representación de los mapas bibliométricos anteriores, se realiza la agrupación utilizando la técnica de clustering por modularidad (ver [Waltman et al., 2010] para información adicional relacionada con esta técnica de agrupamiento). La Figura 2.4, publicada en nuestro artículo en nuestro artículo [Vallez et al., 2021], muestra el grafo de palabras clave mencionado, en el cual los temas están agrupados en 5 categorías:

1. **Verde:** representa investigaciones centradas en el análisis de viajes, implementaciones de BSS con estaciones y el problema de reequilibrio.
2. **Rojo:** agrupa investigaciones relacionadas con conceptos sociales como salud, movilidad, etc.
3. **Azul:** muestra las tendencias más recientes de investigación (BSS sin estaciones o dockless), destacando que este tipo de investigaciones se realizan principalmente en Asia,
4. **Amarillo:** investigaciones realizadas en áreas relacionadas con la economía y el gobierno, destacando que este tipo de trabajos se han llevado a cabo principalmente en América.
5. **Morado:** investigaciones centradas en algoritmos aplicados a los BSS, principalmente utilizando datos europeos.

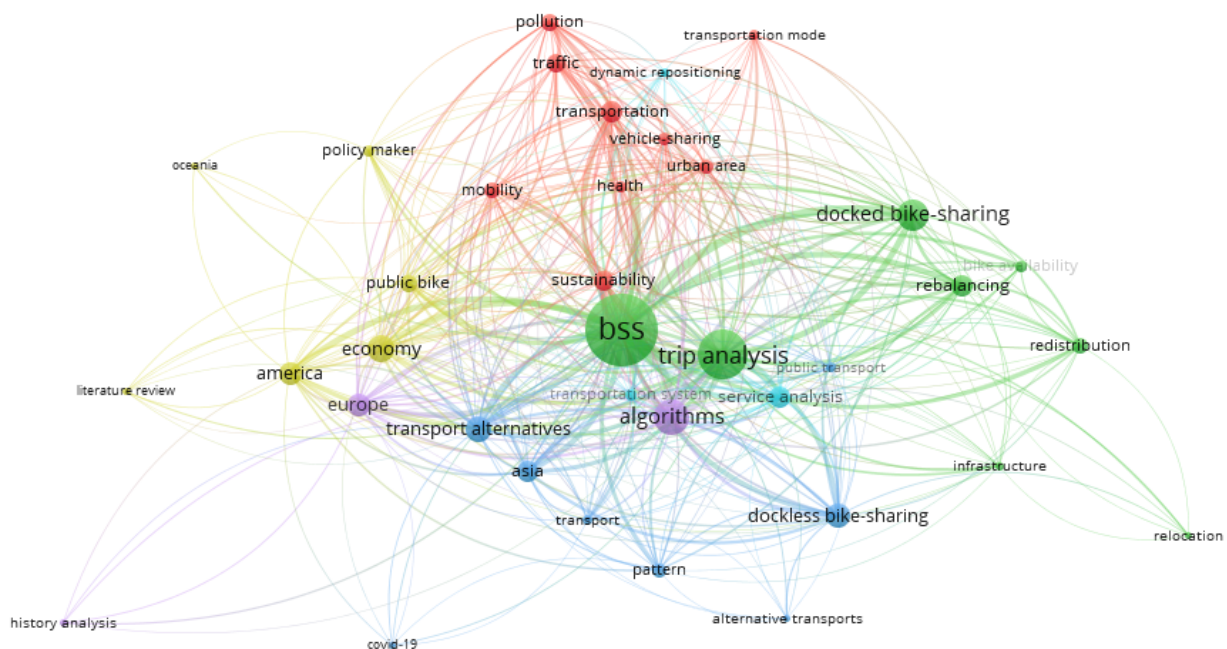


Figura 2.4: Visualización de los conceptos de investigación sobre BSS agrupados por similitud utilizando VOSviewer [Van Eck and Waltman, 2010]. El tamaño es proporcional al número de apariciones del concepto y os diferentes colores representan distintos grupos. Fuente: Elaboración propia publicada en Vallez et al. [2021]

a cabo en Barcelona y [Woodcock et al., 2014] en Londres, que concluyen que la aparición de los BSS pueden producir beneficios para la salud de los usuarios por la reducción de emisiones contaminantes. Sin embargo destacan que aparecen otros riesgos para la salud como son posibles accidentes derivados de la circulación urbana. Posteriormente en [Mueller et al., 2015] se realiza una revisión sistemática de la literatura relacionada con el impacto en la salud de los medios de transporte activos como son las bicicletas. Tiempo después se realiza en [Otero et al., 2018] se realiza un estudio que involucra a doce implementaciones de BSS de Europa. En él se concluye que los BSS constituyen una herramienta tanto para promocionar la salud como para prevenir problemas, y que su impacto depende de las características propias de cada ciudad como son su calidad inicial del aire, la penetración que ha tenido como medio de transporte y la disponibilidad de vías específicas para bicicletas las cuales reducen los siniestros. Destacar, para finalizar esta vertiente de la agrupación social, que en los últimos años se siguen realizando investigaciones en localizaciones concretas como es el caso de [Ou et al., 2023] en China y de [Mora et al., 2024] en Chile, además de contar con un nuevo artículo de revisión del impacto en la salud [Ding et al., 2024] pero con un enfoque generalista, es decir, no centrado exclusivamente en BSS.

En lo referente a la segunda vertiente de esta agrupación, los estudios sobre patrones de viaje comienzan con [Zhou, 2015]. En el artículo [Bordagaray et al., 2016] constatan la heterogeneidad de los patrones. Numerosos autores como [El-Assi et al., 2017; Chen et al., 2018; Du and Cheng, 2018; Kou and Cai, 2019; Eren and Uz, 2020] identifican varios factores que pueden influir en la forma que tienen los usuarios de viajar entre los cuales se encuentra la climatología. [Godichon-Baggioni et al., 2019] señala la importancia de disponer datos históricos para poder realizar estudios sobre los patrones de uso. Finalmente y respecto a la barreras que afecta al servicio destacan principalmente el uso del coche, clima, precio y la cercanía de estaciones a lugares de interés como pueden ser el hogar o el centro de trabajo. Diversos estudios [Fishman et al., 2012; McNeil et al., 2018; Bakogiannis et al., 2019] han abordado la implicación de estas barreras en los BSS

3. Cluster 3 de color morado: El tercer grupo está centrado en el concepto de la demanda del servicio por parte de los usuarios y su impacto en la optimización del servicio. Inicialmente las investigaciones se centraron en la percepción subjetiva que tenían los usuarios del servicio y la influencia que esta percepción ejerce sobre la demanda. Por tanto se centraron en evaluar la heterogeneidad del uso del servicio desde diferentes perspectivas como son la diferenciación entre usuarios frecuentes y ocasionales [Shaheen et al., 2010; Ricci, 2015], los rangos de edad [Shaheen et al., 2011] y si eran usuarios antiguos o recientes [Fishman et al., 2013] entre otros. [Barbour et al., 2019] constató que factores como el género, la edad, los ingresos, la tipología del viaje, la duración del mismo y la disponibilidad de vehículos alternativos influían en el uso de los BSS y en la sustitución de medios de transporte alternativos por éstos. Se puede decir que los estudios durante un tiempo se centraron por tanto en la recopilación de información acerca de los usuarios bien sea por datos proporcionados por el proveedor DeMaio [2009]; Shaheen et al. [2010]; Guo et al. [2017] del servicio o por encuestas realizadas a los usuarios y su posterior análisis multivariante [Godichon-Baggioni et al., 2019]. En estudios posteriores se empieza a modelar la demanda mediante el uso de diferentes algoritmos, partiendo de las regresiones más sencillas, modelos de solver, hasta los algoritmos de aprendizaje automático más avanzados. Si bien la literatura comienza a mezclar el análisis de la demanda con la resolución del problema de rebalanceo, y por tanto, con el fin de evitar duplicidades el detalle de estos estudios se aborda en la sección 2.5.1. Adicionalmente [Eren and Uz, 2020] aporta una revisión de los factores que afecta a la demanda
4. Cluster 4 de color azul: Este conjunto de investigaciones está más relacionado con el ámbito económi-

co, modelos de negocios, gestión del servicio, políticas de sostenibilidad e innovación.

Varios autores han realizado análisis comparativos entre las distintas implementaciones en diferentes localizaciones [Shaheen et al., 2013; Bouveyron et al., 2015; Ricci, 2015; Zhang et al., 2015; Li et al., 2019a; Galatoulas et al., 2020; Shin, 2021; Mahajan and Argota Sánchez-Vaquerizo, 2024; Inhof et al., 2024]. Por otra parte otros describieron y evaluaron distintos modelos de negocio [van Waes et al., 2018a; Curtis and Mont, 2020; Shao et al., 2020; Si et al., 2021; Gao et al., 2024]

Entre las principales conclusiones obtenidas los destaca que los BSS tienen diversos grados de éxito dependiendo de su localización y los apoyos gubernamentales y empresariales. Es decir, están muy ligados a las características socioeconómicas locales.

[van Waes et al., 2018b] comentó que los modelos de negocio basados en estaciones están bien establecidos pero son difíciles de escalar. Sugirió que habría que ir hacia modelos tipo “*free-floating*” ‘por aportar mayor potencial de escalabilidad. Todo esto englobado dentro del ámbito de la innovación.

Finalmente comentar que en este grupo también tienen cabida simuladores relacionados con las políticas del servicio como los descritos por [Chen et al., 2018]. En otros artículos como [Lu et al., 2018; Guo et al., 2019] los autores exploran la simulación basada en agentes.

5. Cluster 5 de color amarillo: La última agrupación de artículos está relacionada con el entorno construido y la sociedad. Indudablemente los BSS ofrecen numerosas ventajas y además en diversos planos frente a otros medios de transporte. Sin embargo para poder promover su uso en la sociedad hay que analizar el entorno donde se implementan y qué factores del mismo influyen. Algunos autores, como [Buck and Buehler, 2012; Singleton and Clifton, 2014; Mattson and Godavarthy, 2017; Wu et al., 2019; Alcorn and Jiao, 2023], se centraron en ese aspecto. Por entorno construido hay que entender las diferencias culturales que influyen, en este caso, en la movilidad.

Los autores de [Guo et al., 2022] realizaron una revisión de las investigaciones realizadas que guardan relación con esta agrupación.

Finalmente comentar que en los últimos años las investigaciones de esta línea se han centrado en implementaciones particulares como es el caso de [Mohiuddin et al., 2023] en Sacramento, [Troped et al., 2023] en Boston, [Wang et al., 2024] en Chicago o [Gong et al., 2024] en Nueva York entre otros.

2.5. El problema del rebalanceo

El problema del rebalanceo es un problema inherente a todos los sistemas de vehículos compartidos VSS que se basan en estaciones donde se puede recoger los vehículos al inicio del viaje y en donde hay que dejarlos al finalizarlo. En el contexto particular de los BSS, el problema surge principalmente en los que están basados en estaciones *Dock-based*. Se debe a que la demanda no está equilibrada entre estaciones y, dentro de una misma estación, es asimétrica (no necesariamente se retiran las mismas bicicletas que se depositan) [DeMaio, 2009; Shaheen et al., 2010]. Cabe mencionar que, ya que está fuera del alcance de esta tesis, en los BSS que no tienen estaciones base (*Free-floating*) existe un problema similar aunque más complejo al no contar con localizaciones concretas y fijas dónde tomar una bicicleta y donde dejarla. La forma de dar solución a esta problemática consiste en gestionar la necesidad de redistribuir las bicicletas entre las distintas estaciones para satisfacer la demanda de los usuarios [Raviv et al., 2013].

El problema de rebalanceo en los sistemas de bicicletas compartidas (BSS) puede clasificarse según dos criterios principales. El primero de ellos se enfoca en la dimensión temporal. En este sentido, se distinguen dos modalidades: el rebalanceo estático y el dinámico. El rebalanceo estático se basa en la utilización de datos

históricos para planificar la redistribución, normalmente llevada a cabo en horarios de menor actividad, como durante la noche [Raviv et al., 2013]. Por el contrario, el rebalanceo dinámico opera en tiempo real y se adapta de manera continua a las condiciones cambiantes del sistema, haciendo ajustes constantes en función de la evolución de la demanda y la disponibilidad de bicicletas [Contardo et al., 2012].

El segundo criterio de clasificación atiende a quién es responsable de la redistribución de las bicicletas. Por un lado, la empresa operadora del servicio puede asumir la tarea de rebalanceo utilizando vehículos —como camiones— para mover las bicicletas de unas estaciones a otras [Benchimol et al., 2011]. Por otro lado, es posible responsabilizar al propio usuario a través de incentivos, de modo que las personas usuarias se vean motivadas a devolver las bicicletas en determinadas estaciones que requieran un mayor equilibrio en ese momento [Fricker and Gast, 2016].

Es importante destacar que en la literatura, dicho problema ha recibido múltiples denominaciones como se discute en la referencia [De Chardon et al., 2016], motivado principalmente porque el problema ha sido estudiado desde diversas disciplinas como son la Economía, la Logística, la Computación, las Matemáticas o la Ingeniería. Esta diversidad terminológica refleja la fragmentación conceptual en este campo de investigación, lo que dificulta la unificación del conocimiento. Para abordar este problema, tal y como se ha comentado previamente, en nuestro primer artículo se creó un tesoro terminológico que organiza el vocabulario. El fin es facilitar la búsqueda exhaustiva en la revisión del estado del arte, permitiendo identificar términos equivalentes, evitando la exclusión de trabajos relevantes simplemente por tener una nomenclatura de problema diferente.

En la Tabla 2.2 se recogen las variantes de denominaciones principales y un listado de referencias que la emplearon.

Variante de denominación	Estudios que usan dicha nomenclatura
balancing	[Benchimol et al., 2011; Rainer-Harbach et al., 2013a; Kloimüllner et al., 2014]
rebalancing	[Erdoğan et al., 2015; Regue and Recker, 2014a]
repositioning	[Raviv et al., 2013; Forma et al., 2015; Han et al., 2016]
relocation	[Erdoğan et al., 2014]
redistribution	[Labadi et al., 2012; Lin and Chou, 2012; Nair et al., 2013; Pfrommer et al., 2014]

Tabla 2.2: Variantes principales empleadas en la literatura para denominar al problema del rebalanceo

El problema del rebalanceo, como establecimos en el primer artículo que forma parte de esta tesis [Vallez et al., 2021], se divide en dos subproblemas clave. En primer lugar, se encuentra el análisis y modelado de la demanda del servicio. Se trata de un desafío complejo debido a la interacción de múltiples variables que impactan de manera desigual en las distintas implementaciones de BSS. En segundo lugar, está la optimización del proceso de redistribución de las bicicletas. En este segundo subproblema hay que conseguir mantener un equilibrio entre los niveles de servicio esperados y los costes asociados. La complejidad de este problema, como establece [De Chardon et al., 2016] se debe a la interacción de múltiples características, limitaciones y restricciones, entre las cuales se destacan las siguientes:

- Limitación de capacidades: número de bicicletas que pueden moverse de una estación a otra, capacidad del vehículo que las transporta, capacidad de las estaciones y recursos operativos disponibles como número de camiones para transportar bicicletas.

- Naturaleza dinámica del problema: las condiciones pueden cambiar en tiempo real obligando a replantearse las acciones en curso.
- Concurrencia de varios actores que influyen directamente en el sistema.
- Restricciones temporales y espaciales: como por ejemplo volumen de tráfico, distancias, velocidad de desplazamiento, hora o tipo de día.
- Necesidad de poder predecir la demanda para cuantificar, en este caso, el número de bicicletas que hay que mover de una estación a otra.

Este segundo subproblema ha sido ampliamente abordado en la literatura científica, dentro de diversas áreas de conocimiento y aplicando diversas técnicas como se recoge en [Caceres-Cruz et al., 2014; Golden et al., 2023; Liu et al., 2023]. En la literatura, este subproblema recibe diferentes nombres dependiendo del enfoque adoptado. Se detallan a continuación los tres más significativos con la referencia original que estableció el problema. Uno de ellos es el *Vehicle Routing Problem (VRP)* [Dantzig and Ramser, 1959], que representa el marco más general del problema, en el cual múltiples vehículos deben atender una serie de clientes de forma eficiente. Otra variante es el *Pickup and Delivery Problem (PDP)* [Savelsbergh and Sol, 1995], una subvariante del VRP que se centra específicamente en la lógica de recoger bienes en un lugar y entregarlos en otro. Finalmente, en su forma más básica, se encuentra el *Travelling Salesman Problem (TSP)* [Applegate, 2006], donde se asume un único vehículo o “vendedor” (como enuncia el problema original) que debe visitar un conjunto de ubicaciones optimizando la distancia recorrida.

La solución a este subproblema requiere de un enfoque integral, que tenga en cuenta todas las restricciones y la interacción entre las mismas ya que no operan de forma aislada.

A pesar de la relevancia del segundo subproblema relacionado con la optimización de la redistribución, esta tesis se enfoca en el primer subproblema, el análisis y modelado de la demanda del servicio, dado que, entre otros aspectos, éste último representa una de las restricciones clave para abordar el segundo.

2.5.1. Cronología de la revisión de la literatura acerca del problema del rebalanceo

Se incluye en esta sección una descripción narrativa de los principales artículos científicos, ordenados cronológicamente, que han estudiado el problema del rebalanceo en los sistemas de bicicletas compartidas. Hasta 2020 dicha información está recogida y publicada en nuestro primer artículo [Vallez et al., 2021], siendo actualizada en este capítulo de la tesis. Para tener un resumen y una visión global de las aportaciones se puede consultar la Tabla E.1 del Apéndice E.

Como se ha comentado, el problema de balanceo ha sido identificado como uno de los principales desafíos que las implementaciones de los BSS han tenido que abordar desde sus primeras implementaciones. Los operadores y gestores del servicio de los BSS reconocieron la necesidad de mejorar los métodos de distribución de bicicletas y, en algunos casos, también evaluaron los costos asociados [DeMaio, 2009]. Por lo tanto, desde los inicios ha existido la necesidad de encontrar métodos que hagan más eficiente la redistribución de las bicicletas [Shaheen et al., 2010].

A lo largo del tiempo se han propuesto diversos enfoques en la literatura para abordar dicho problema.

En el artículo [Nair and Miller-Hooks, 2011], utilizando programación estocástica, asumieron que el coste de mover vehículos (coches o bicicletas) puede representarse mediante una función fija más lineal, sin tener en cuenta las restricciones y costes de las rutas. El modelo asume una demanda estática y no considera fluctuaciones en la demanda en diferentes periodos, por lo que este modelo se considera demasiado simplista para el problema de reposicionamiento de los BSS.

Otro enfoque inicial [Benchimol et al., 2011] describe el problema como una alternativa al clásico problema del “*Travelling Salesman Problem*” (*TSP*) y propone una metodología basada en algoritmos de aproximación para reducir el coste. Ese mismo año, otro estudio [Wood et al., 2011] utilizó técnicas de visualización de origen-destino en estaciones para mostrar los flujos de viajes entre estaciones de anclaje. Además de identificar posibles estrategias de reequilibrio, la principal contribución de este artículo fue reconocer la necesidad de un estudio de demanda para encontrar una estrategia óptima de balanceo. De ahí se deriva la necesidad. Además, encontraron una clara dependencia entre la estrategia de reequilibrio y la implementación concreta del BSS, lo que también se evidenció en [Borgnat et al., 2011].

En el trabajo [Chemla et al., 2012] introdujeron un marco teórico más riguroso para comprender mejor los factores críticos que afectan la solución óptima. Además, este fue el primer caso en el que se utilizó un enfoque metaheurístico en este contexto. Propusieron una solución basada en programación entera con varias restricciones resueltas mediante un algoritmo de ramificación y corte (“*branch-and-cut*”).

En el año 2012, en [Labadi et al., 2012] propusieron el uso de redes de Petri para crear un modelo dinámico modular. El objetivo era buscar valores óptimos que minimizaran los escenarios de estaciones vacías y llenas. Por otra parte [Lin and Chou, 2012] reconocieron que la redistribución de bicicletas es parte del problema de enrutamiento de vehículos (*VRP*) y que tradicionalmente se resolvía usando la distancia Euclidiana con el fin de calcular el costo de transporte. Adicionalmente propusieron también incluir información geográfica, como condiciones de las carreteras y regulaciones de tráfico proporcionada por sistemas de información geográfica (*GIS*), en el cálculo del coste de transporte. Con este enfoque y probando diferentes métodos de construcción, concluyeron que la inserción más lejana (*farthest insertion*) funciona mejor que el enfoque del vecino más cercano (*nearest-neighbor*).

Hasta 2013 la forma de abordar el problema se basaba en la idea principal de suponer que existía un nivel de ocupación óptimo por estación y que el objetivo es lograr esos niveles con el costo mínimo de reposicionamiento de bicicletas. [Raviv et al., 2013] ofrecieron dos formulaciones del problema. La primera utilizando programación lineal entera mixta (*MILP*, por sus siglas en inglés, “*Mixed-Integer Linear Programming*”), obteniendo soluciones con brechas de optimalidad aceptables para implementaciones de BSS de tamaño moderado (hasta 60 estaciones). La segunda formulación es para sistemas más grandes. En este caso sugirieron aplicar su método junto con un enfoque geográfico o usar métodos heurísticos como la búsqueda tabú (*Tabu search*) o algoritmos genéticos.

Otro enfoque para resolver el problema de balanceo fue mediante el uso del algoritmo de la búsqueda de vecindad variable (*VNS*, por sus siglas en inglés, “*Variable Neighborhood Search*”) [Raidl et al., 2013; Rainer-Harbach et al., 2013b]. En general, *VNS* tiene un buen rendimiento y escala mucho mejor que los enfoques de programación lineal. Los autores de [Nair et al., 2013] decidieron obtener un análisis cuantitativo de una implementación de gran escala de un BSS (sistema Vélib’ en París). Analizaron el problema de redistribución utilizando información histórica estocástica. Esta investigación se basa en un estudio previo del mismo autor principal [Nair and Miller-Hooks, 2011]. Además, ese mismo año, [Lin et al., 2013] propusieron un diseño estratégico para los BSS teniendo en cuenta las consideraciones de inventario y formulando el problema como un modelo de localización de hubs con inventario (*Hub location inventory model*). Usaron heurísticas para calcular el costo de la función objetivo. Su punto débil es que no consideran en su modelo que las bicicletas registradas pueden reutilizarse.

Alternativamente, en [Di Gaspero et al., 2013] abordaron el problema utilizando la llamada programación por restricciones. Para acelerar la búsqueda de las rutas más prometedoras, utilizaron un enfoque de búsqueda en vecindades amplias (*LNS*, por sus siglas en inglés, “*Large Neighborhood Search*”). En 2014, en el trabajo [Erdoğan et al., 2014] introdujeron, modelaron y resolvieron una variante del problema del Viajante de Comercio (*TSP*) con Recogida y Entrega de una Única Mercancía (*1-PDTSP*, por sus siglas en inglés, “*One Commodity Pickup and Delivery Traveling Salesman Problem*”). La denominaron Problema de Reubi-

cación Estática de Bicicletas con Intervalos de Demanda (*SBRP-DI*, por sus siglas en inglés, “*Static Bicycle Relocation Problem with Demand Intervals*”). Este problema consiste en reubicar bicicletas entre diferentes estaciones considerando el inventario inicial de cada estación y algunos límites inferiores y superiores fijos. Lo resolvieron aplicando programación entera con un algoritmo de ramificación y corte. Ese mismo año, [Kloimüller et al., 2014] ampliaron su trabajo previo realizado desde un punto de vista estático (usando heurísticas Greedy-PILOT con búsqueda de vecindad variable). Otros autores [Regue and Recker, 2014b] propusieron un marco para resolver el problema de reposicionamiento basado en cuatro modelos dentro del paradigma de programación dinámica:

1. Pronóstico de demanda
2. Inventario de estaciones
3. Necesidades de redistribución
4. Enrutamiento de vehículos

Desde la perspectiva de costo-beneficio, en [Pfrommer et al., 2014][42] los autores combinaron el método clásico de redistribución realizado por el propietario de la implementación del BSS con un enfoque de incentivos de precios a los clientes. Este fue el primer ejemplo de un enfoque de rebalanceo basado en usuarios mezclado con uno clásico basado en operadores, demostrando que se pueden lograr reducciones de costos haciendo cambios dinámicos en las recompensas otorgadas a los usuarios si estacionan en una u otra estación de bicicletas.

En el estudio [Dell’Amico et al., 2014] abordaron el problema combinando cuatro métodos diferentes y comparando sus resultados. Todas las alternativas tenían en común un número exponencial de restricciones, por lo que adoptaron un algoritmo de ramificación y corte para resolverlas. Tras probarlo con varios conjuntos de datos, concluyeron que este enfoque funciona bien para instancias de hasta 50 vértices. Para problemas más grandes, el tiempo de cálculo aumenta, por lo que sugieren tratarlos con algoritmos heurísticos y metaheurísticos.

Ese mismo año, se introdujo un método para identificar los patrones de viaje de los usuarios mediante el análisis de los comportamientos de desplazamiento [Beecham et al., 2014]. Según esta investigación, el análisis clásico de la demanda en horas pico causó más problemas al reequilibrio que las técnicas de clasificación visual utilizadas en este trabajo (cálculo de centros de gravedad ponderados, agrupamiento k-means espacial y estimación de densidad de kernel). Con este enfoque, se identificaron las principales fuentes de desequilibrio (mañanas frente a tardes o viajes originados dentro de Londres frente a los originados fuera de ella, entre otros).

[Ho, 2014] utilizó una heurística de búsqueda tabú iterativa que produjo soluciones de alta calidad con tiempos de cómputo reducidos. Además de la solución mejorada, este artículo contribuyó a aumentar el realismo y a reducir el espacio de soluciones del problema de reequilibrio. Finalmente, los autores comentaron que este enfoque podría usarse con otras funciones objetivo y recomendaron extender la investigación a escenarios con múltiples vehículos de reposicionamiento.

Finalmente, también en 2014, [Zhao et al., 2014] analizaron el grado de éxito de los sistemas de bicicletas compartidas (BSS) en China. Aunque este trabajo no presentó algoritmos ni soluciones concretas al problema de reequilibrio, es importante señalar que los autores concluyeron que la estrategia de reequilibrio adoptada influye en el nivel de servicio, siendo este un factor clave frecuentemente ignorado para lograr una implementación exitosa de BSS.

A principios de 2015, se propuso un enfoque heurístico de tres pasos [Forma et al., 2015]. Primero, dividir y agrupar las estaciones considerando factores geográficos y de inventario (agrupamiento). En segundo

lugar, se establecen las rutas de los vehículos a través de los grupos tomando decisiones sobre el nivel de inventario para cada estación. Por último, el tercer paso consiste en permitir transferencias únicamente entre estaciones del mismo grupo o entre grupos consecutivos. Ese mismo año, [Erdoan et al., 2015] introdujeron un algoritmo de programación lineal, convirtiendo las variables generales en binarias. Esta solución fue probada con varios conjuntos de datos, logrando escalar el cálculo hasta 60 estaciones.

En 2016, [Alvarez-Valdes et al., 2016] discutieron el papel del plan de reposicionamiento como un factor clave en la calidad del servicio. Proporcionaron un resumen de la investigación realizada hasta entonces y concluyeron que, aunque cada sistema tiene sus propias características, todos comparten dos partes fundamentales. La primera es la predicción/pronóstico de la demanda estocástica y la segunda es el plan de enrutamiento. Su contribución consistió en diseñar un procedimiento que combina ambas partes. Finalmente, recomendaron extender la investigación a soluciones dinámicas. En [Szeto et al., 2016] replantearon el problema y lo analizaron como un algoritmo de optimización por reacciones químicas (*CRO*, por sus siglas en inglés, “*Chemical reaction optimization*”), obteniendo mejores soluciones con tiempos de cómputo reducidos.

En cuanto a los tiempos de cómputo, en [Dell’Amico et al., 2016] afirmaron que su solución propuesta mejoró significativamente las alternativas anteriores al combinar una nueva heurística con búsquedas locales y adaptarla mediante una técnica de ramificación y corte para considerar el tiempo máximo de duración del reposicionamiento.

[Li et al., 2016] abordaron el escenario de tener diferentes tipos de bicicletas. Propusieron reducir el costo total de la tarea de reequilibrio formulando un problema de programación lineal mixta y resolviéndolo con un algoritmo genético híbrido.

En otra investigación interesante publicada en 2016 por [Médard de Chardon et al., 2016], se describió el proceso de reequilibrio de nueve sistemas de bicicletas compartidas. Concluyeron que la relación entre los viajes y el reequilibrio es más compleja de lo esperado. Aunque no presentaron un nuevo algoritmo de reequilibrio, una de sus conclusiones es altamente relevante: el simple hecho de aplicar una política de reequilibrio en una estación afecta directamente su nivel de demanda. Otros autores, como [Cruz et al., 2016], se centraron en una versión más simple del problema (utilizando solo un vehículo para reequilibrar bicicletas entre estaciones). El algoritmo propuesto en esta investigación es una heurística basada en búsqueda local iterativa (*ILS*, por sus siglas en inglés, “*Iterated local search*”), y también realizaron una comparación con soluciones propuestas anteriormente [Chemla et al., 2012; Erdoan et al., 2015].

En la investigación [Liu et al., 2016] destacaron la división del problema de reequilibrio en dos partes: la primera relacionada con la decisión del nivel objetivo de inventario y la segunda sobre el plan de enrutamiento para distribuir bicicletas con el fin de alcanzar dicho nivel objetivo. Utilizaron un regresor *MSWK* (“*Meteorology Similarity-Weighted K-Nearest-Neighbor*”) para predecir la demanda basada en datos históricos y un modelo de programación no lineal entera mixta (*MINLP*, por sus siglas en inglés, “*Mixed-Integer Non-Linear Programming*”) para minimizar la distancia recorrida por el vehículo de enrutamiento. Finalmente, para agrupar las estaciones, los autores usaron un algoritmo de Agrupamiento de K-Centros con Restricciones de Capacidad Adaptativa (*AdaCCKC*, por sus siglas en inglés, “*Adaptive-Capacity Constrained K-centers Clustering*”).

Asimismo, en 2016, [Zhang, 2017] propusieron una metodología dinámica de reposicionamiento de bicicletas. Con el pronóstico del nivel de inventario, la llegada de usuarios y datos de reposicionamiento de bicicletas, crearon un modelo de red espacio-tiempo traducido en un complejo problema de optimización no lineal, analizado con un algoritmo heurístico personalizado.

En esta línea, [Ho and Szeto, 2017] propusieron un algoritmo heurístico basado en una búsqueda en vecindades amplias (*LNS*), con búsqueda tabú y varios operadores de inserción y eliminación para mejorar el rendimiento algorítmico. Este algoritmo supera a los anteriores, pero tiene una limitación: la posibilidad de

mover bicicletas desde estaciones ya equilibradas hacia estaciones vecinas para reducir costos de transporte.

De manera similar, en [Shui and Szeto, 2017a], utilizando el algoritmo de colonia artificial de abejas mejorada (*EABC*, por sus siglas en inglés, “*Enhanced Artificial Bee Colony*”), se centraron en reducir las emisiones de CO_2 , superando al clásico algoritmo genético.

[Chiariotti et al., 2017] emplearon una combinación de procesos de nacimiento-muerte para decidir el momento de redistribuir bicicletas y teoría de grafos para seleccionar la ruta de reequilibrio.

En 2017, [Elhenawy and Rakha, 2017] propusieron una solución basada en la teoría de juegos, específicamente en el algoritmo de aceptación diferida. Este algoritmo tiene dos fases: construir la ruta de reequilibrio y mejorar dicha ruta. Al final, el problema se modela como dos conjuntos disjuntos de jugadores, cada uno con sus propias decisiones.

Finalmente, en 2017, [Schuijbroek et al., 2017] combinaron dos aspectos que tradicionalmente se manejaban por separado: el nivel de servicio de cada estación de bicicletas y la optimización de las rutas de los vehículos para reequilibrar bicicletas. Para combinar ambos subproblemas, primero utilizaron un algoritmo de agrupamiento y, posteriormente, aplicaron algunas heurísticas.

En [Bulhões et al., 2018] incluyeron la posibilidad de utilizar múltiples vehículos con varias visitas para ejecutar el reequilibrio entre estaciones. Propusieron un problema de programación entera, resuelto mediante un algoritmo de ramificación y corte. Su principal contribución fue demostrar que múltiples visitas solo resultan interesantes para vehículos con una capacidad superior a 20 bicicletas. También concluyeron que el número de visitas necesarias disminuye con el aumento en el número de vehículos disponibles.

[Shui and Szeto, 2017b] propusieron un algoritmo de colonia artificial de abejas mejorada (*EABC*) concluyendo que (*EABC*) tiene un mejor desempeño y ofrece soluciones más aproximadas que el algoritmo clásico de CPLEX. También proporcionaron una lista de recomendaciones relacionadas con la duración de las rutas, el tiempo de servicio, el número de vehículos y la satisfacción de la demanda, todas ellas respaldadas por evidencia. [Legros, 2019] desarrolló una herramienta de soporte de decisiones basada en un proceso de decisión de Markov, que permite al operador decidir en un momento específico cuántas bicicletas deben moverse de o hacia cada estación. Por otro lado, [Dell’Amico et al., 2018] ofrecieron una solución heurística al problema, enfocándose en una versión con demanda estocástica.

Paralelamente, [Hulot et al., 2018] contribuyó a la solución del problema mejorando la predicción de la demanda mediante técnicas de aprendizaje automático.

Otro ejemplo de un modelo dinámico de reposicionamiento de bicicletas fue presentado por [Li et al., 2018], quienes utilizaron un aprendizaje por refuerzo espaciotemporal basado en algoritmos de agrupamiento. Dividieron el sistema de bicicletas compartidas en varios grupos y aplicaron el modelo a cada grupo. De esta forma, redujeron la complejidad del problema y lograron aprender políticas óptimas de reposicionamiento entrenando el modelo.

En 2019, [You, 2019] propuso un modelo de programación no lineal mixta con restricciones. Para abordar la complejidad del problema, lo dividió en subproblemas independientes del tiempo. Luego, resolvió los problemas restantes utilizando modelos de programación lineal seleccionados. La evidencia muestra que este enfoque es superior a GAMS/CEPLEX y también a un enfoque mixto (algoritmo genético más programación lineal).

[Tang et al., 2019] resolvieron un modelo de dos niveles utilizando algoritmos de búsqueda local iterativa y búsqueda tabú. El nivel superior gestiona la cantidad de bicicletas que deben moverse entre las estaciones objetivo, mientras que el nivel inferior se centra en reducir los costos de transporte. Según sus pruebas, su rendimiento es mejor que las soluciones disponibles anteriormente. Además, proporcionaron algunas ideas sobre los factores que afectan el costo de reposicionamiento: el tamaño del vehículo de reposicionamiento, el valor del tiempo de carga/descarga de una sola bicicleta y la duración del tiempo de reposicionamiento.

[Wang and Wu, 2019] incluyeron variables adicionales en el problema (bicicletas y estaciones inutiliza-

bles, múltiples vehículos y visitas, y restricciones debidas a los requisitos de reparación de estaciones). Propusieron un modelo probado con un problema pequeño y resuelto con algoritmos matemáticos heurísticos simples. Concluyeron que para investigaciones más amplias y grandes, se requiere el uso de algoritmos metaheurísticos como la búsqueda local iterativa, la búsqueda en vecindad amplia o la búsqueda tabú.

[Warrington and Ruchti, 2019] propusieron una aproximación estocástica de dos etapas aplicada no solo a BSS, sino también a otros sistemas de vehículos compartidos. Se incluyó en esta revisión porque utilizaron un sistema de BSS para evaluar su solución.

Durante 2020, muchos trabajos se enfocaron en anticipar la demanda futura. En particular, [Brinkmann et al., 2020] introdujeron un problema de enrutamiento de inventario dinámico-estocástico modelado como un proceso de decisión de Markov para resolver el problema. Además, [Tang et al., 2020] centraron su investigación en minimizar varios costos asociados con el procedimiento de reequilibrio: 1. costo de transporte, 2. costo por penalización en las estaciones al no satisfacer la demanda y 3. costos de almacenamiento. Para ello, utilizaron un modelo de programación estocástica de dos etapas resuelto con un algoritmo de recocido simulado (“*simulated annealing algorithm*”).

Paralelamente, [Maleki et al., 2020] modelaron el problema como una red de Jackson (modelo matemático para análisis de colas) y utilizaron un algoritmo genético para resolverlo. Obtuvieron mejoras aceptables y también proporcionaron recomendaciones sobre el tamaño de la flota y la capacidad del sistema BSS.

Otra solución fue propuesta por [Zhang et al., 2020], quienes modelaron el problema como uno de programación lineal clásica, pero lo resolvieron con un algoritmo de búsqueda tabú adaptativa (*ATS*, por sus siglas en inglés, “*Adaptive Tabu Search*”).

[Jia et al., 2020], en lugar de centrarse en obtener el punto de equilibrio como la literatura clásica, introdujeron el concepto de “intervalo de equilibrio”. Lo modelaron como un problema de programación entera mixta biobjetivo, utilizando un algoritmo de optimización por enjambre de partículas multiobjetivo y de múltiples comienzos (MS-MOPSO).

Finalmente, Lu et al. [2020] propusieron un algoritmo memético que combina un método de construcción voraz para generar soluciones, un algoritmo de copia de rutas para la recombinación de soluciones y una búsqueda local para la mejora de soluciones. Esta solución proporcionó una doble mejora: primero, el 46 % de los casos obtuvieron un mejor resultado conocido; y segundo, fue de dos a seis veces más rápido que las soluciones disponibles en la literatura.

Con respecto al año 2021 destacamos inicialmente nuestro paper [Vallez et al., 2021] el cual incluye entre otras cosas una revisión de los papers hasta 2020, los cuales se han comentado anteriormente hasta este punto.

2.6. Conclusiones del capítulo

El análisis desarrollado en este capítulo permite construir una visión integral sobre el estado del arte en el ámbito de los BSS. Se recopilan las investigaciones más relevantes, se definen las tendencias temáticas temporales y se identificaron desafíos pendientes aún por abordar. Entre los desafíos iniciales destacados se encuentra la fragmentación terminológica. Se trata de un problema que ha estado presente en la literatura desde los primeros estudios. Esta cuestión se aborda en el primer artículo de esta tesis citeVallez2021, donde se establece una metodología que garantizara la recopilación adecuada de los artículos disponibles.

Otro aspecto relevante del estado del arte se refiere a la diversidad de disciplinas en las cuales se han realizado investigaciones en este campo de conocimiento. Se encuentran estudios destacados relacionados principalmente con los ámbitos económico, social, sanitario, de planificación urbanística, sostenibilidad y optimización logística. Todo esto generó una necesidad de clusterización temática de la literatura, la cual fue

inicialmente abordada en nuestro primer artículo. Posteriormente, esta clasificación se actualizó y replicó en otros artículos de revisión. Seguidamente se identifica el problema del rebalanceo de bicicletas como el principal desafío dentro del ámbito de los *BSS*. Este hecho motivó que este estado del arte se centrara en dicha problemática. El problema del rebalanceo está compuesto por dos subproblemas: el primero consiste en la predicción de la demanda de bicicletas, y el segundo se centra en la optimización del traslado de bicicletas entre estaciones con el fin de satisfacer dicha demanda. A continuación se trazó la evolución histórica del uso de distintas herramientas para abordar el problema del rebalanceo. Desde los primeros enfoques estocásticos y heurísticos hasta la incorporación más reciente de algoritmos de aprendizaje automático avanzados como los categorizados como *deep learning*. La revisión y categorización de los algoritmos de aprendizaje automático reveló una clara polarización en la literatura hacia el uso de algoritmos deterministas frente a los probabilísticos. Esta observación condujo a la identificación de una nueva brecha, abordada en el segundo artículo de esta tesis [Vallez et al., 2024]. Se realiza una implementación de un algoritmo de aprendizaje automático probabilístico. En resumen, el análisis presentado en este capítulo responde a la pregunta de investigación RQ2 y sus correspondientes subpreguntas descritas en el Capítulo 1.5.2, cumpliendo también con los objetivos establecidos OB2.1 y OB2.2 detallado en el Capítulo 1.5.3.

“The more you know about the past, the better prepared you are for the future.”

Cuanto más sepas sobre el pasado, mejor preparado estarás para el futuro.

— Theodore Roosevelt

3

Análisis exploratorio de los datos (EDA)

Índice del capítulo

3.1. Definición del alcance de los datos empleados	54
3.2. Descripción y procedencia de los datasets iniciales	55
3.3. Extracción y transformación de los datos	64
3.4. Análisis exploratorio de variables relevantes	73
3.5. Resultados y conclusiones	75

En este capítulo se describe detalladamente el proceso de extracción de los datos empleados en esta investigación. Abarca desde la identificación de las fuentes de información y la obtención de archivos estructurados para su procesamiento, hasta el enriquecimiento de los datos mediante fuentes complementarias. Adicionalmente, se delimita y justifica el alcance del análisis, considerando las restricciones, desafíos y eventos extraordinarios de los años más recientes. Como cierre, se realiza un análisis exploratorio de los datos seleccionados y se presenta el conjunto de datos que se empleará en los próximos capítulos para la implementación y validación del modelo realizado en esta tesis.

Las preguntas a las cuales da respuesta este capítulo son: *¿Cuáles son los patrones de uso del servicio BiciMad considerando el perfil de los usuarios, los horarios de utilización, la duración de los viajes y las distancias recorridas? ¿Cuáles son las variables esenciales identificadas en los datos de viajes que permiten modelar la demanda del servicio? ¿Es necesario incluir datos adicionales de otras fuentes o variables nuevas como resultado de la combinación y transformación de las existentes?*

3.1. Definición del alcance de los datos empleados

Los datos que se van a extraer, transformar y analizar se limitan a los correspondientes a los viajes realizados en BiciMad durante los años 2018 y 2019. Estos años fueron seleccionados por su estabilidad operativa y su capacidad para representar el funcionamiento habitual del sistema. La decisión de no incluir períodos

posteriores responde al hecho de encontrar alteraciones significativas en la demanda del servicio, ocasionadas por eventos excepcionales. Así por ejemplo nos encontramos con la incidencia de la pandemia de COVID-19 en el año 2020 con sus períodos de obligado confinamiento y vuelta gradual a la normalidad o el de la tormenta Filomena en 2021 que colapsó la ciudad de Madrid. Ambos eventos supusieron impactos notables en la demanda del servicio, cambiando el comportamiento esperado de los usuarios e incluyendo periodos de indisponibilidad del mismo. Respecto a los años 2022, 2023 y 2024 los datos presentan diversas limitaciones y complicaciones debido a problemas de acceso a la información, modificaciones en la infraestructura y en el sistema tarifario del servicio, incluyendo períodos de gratuidad, y recientes cambios regulatorios municipales que han impactado en el mercado competitivo. Por estas razones se ha decidido excluir del alcance de esta tesis, dejando abierta la posibilidad de retomar su estudio en trabajos futuros, cuando el servicio alcance una mayor estabilidad, permitiendo una comparación robusta con años anteriores. Futuras investigaciones abordarán el análisis de los datos de esos años partiendo del modelo calibrado para 2018 y 2019.

3.2. Descripción y procedencia de los datasets iniciales

Para modelar la demanda de BiciMad, el presente estudio se apoya en cinco conjuntos de datos que proporcionan información relevante en diferentes dimensiones. En particular, comprenden: datos de los viajes registrados en el sistema, parámetros característicos de las estaciones de BiciMad como su localización, registros meteorológicos, un calendario de días festivos y un dataset adicional creado específicamente a partir de información recopilada en tiempo real mediante una API, que detalla el estado operativo de las estaciones.

3.2.1. Dataset 1 - Viajes realizados

Los datos acerca de los viajes realizados en el servicio BiciMad se han obtenido del sitio web de Open Data de la Empresa Municipal de Transportes de Madrid (EMT Madrid) [EMT Madrid, 2024a]. EMT Madrid es una entidad pública que depende del Ayuntamiento de Madrid y se ocupa de gestionar y operar varios servicios de transporte público dentro de la ciudad. urbano en Madrid. Sus principales ámbitos de actuación son la Red de autobuses urbanos, BiciMad, Grúas y Aparcamientos municipales y Teleférico de Madrid. Su misión es la de garantizar una movilidad eficiente, sostenible y accesible para los ciudadanos de Madrid.

La Tabla 3.1 proporciona una visión general de las propiedades cualitativas y cuantitativas de los archivos originales disponibles en la fuente de datos correspondientes al año 2018. Este año se considera representativo de las características generales de los datos utilizados en el análisis. La Figura 3.1 muestra los histogramas ajustados que representan el número de viajes para las distintas distancias. Estos histogramas ya han sido depurados para excluir los viajes asociados a estaciones inauguradas en 2019 y 2020, dado que incluir estas estaciones generaría un sesgo en los datos. Incorporar estaciones nuevas introduce tipos de viajes que no existían en 2018, lo que haría que los resultados no fueran comparables. Además, los histogramas reflejan la significativa disminución de viajes registrada en 2020 debido a los efectos de la COVID-19, incluyendo los períodos de confinamiento en los que los servicios fueron suspendidos. Durante dicho confinamiento, los servicios se interrumpieron completamente y su reactivación se llevó a cabo de manera gradual, tras el retiro y posterior reabastecimiento de bicicletas en las estaciones.

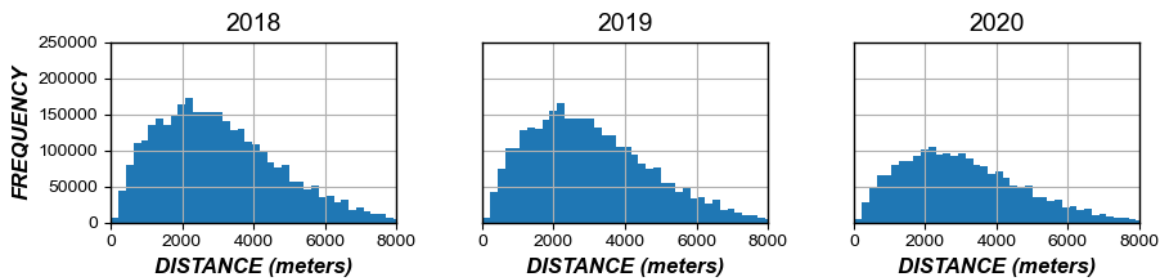


Figura 3.1: Histograma del número de viajes por distancia para los años 2018, 2019 y 2020. Observe cómo el patrón es similar en 2018 y 2019, pero la COVID-19 cambió drásticamente el volumen de viajes durante 2020.

Concepto del Campo	Valor
Datos Disponibles	Desde abril de 2017 hasta junio de 2021
Número de archivos	17 archivos (mensuales, excepto los períodos de julio a diciembre de 2019 y de enero a mayo de 2020, que están integrados en un solo archivo por período).
Formato de los archivos	Archivo comprimido en formato Zip, que contiene un archivo JSON
Tamaño máximo del archivo JSON (dividido por meses)	1.094.440 KB en septiembre de 2018
Tamaño mínimo del archivo JSON (dividido por meses)	3.363 KB en abril de 2020
Tamaño promedio del archivo JSON (dividido por meses)	336.906 KB en 51 archivos

Tabla 3.1: Propiedades cualitativas y cuantitativas de los archivos originales de datos sobre viajes de BiciMad

Como se detalla en el capítulo 3.3, centrado en la descripción de los procesos de extracción y transformación de los datos, el formato original en el cual se suministran estos datos es ficheros JSON. Para una mejor comprensión de la estructura de este fichero, la Tabla 3.2 presenta una descripción de los campos contenidos en el archivo fuente, basada en la documentación oficial del modelo de datos proporcionada por EMT Madrid [EMT Madrid, 2024c]. Además, la Tabla 3.3 muestra un ejemplo representativo de un registro extraído del fichero JSON, mientras que la Imagen 3.2 ilustra gráficamente la estructura jerárquica de este mismo ejemplo.

3. Análisis exploratorio de los datos (EDA)

Nombre del campo	Descripción	Tipo
_id	Este campo contiene a su vez un subcampo denominado \$oid el cual almacena el identificador único del viaje. El formato es formato de ObjectId (común en bases de datos como MongoDB)	Texto
user_day_code	Código único para identificar todos los trayectos de un usuario particular en un día. Puede emplearse para realizar estudios estadísticos de usuarios pero a nivel de día, ya que al día siguiente el código de un mismo usuario cambiará.	Texto
idplug_base	Número del punto de anclaje de la estación de destino.	Entero
user_type	Número que indica el tipo de usuario que realizó el viaje. Sus valores posibles son: 0: No se pudo determinar el tipo de usuario. 1: Usuario anual (titular de un abono anual). 2: Usuario ocasional (perfil turístico). 3: Empleado de empresa.	Entero
idunplug_base	Número del punto de anclaje de la estación de origen.	Entero
travel_time	Tiempo total en segundos entre el desacoplamiento y el reacoplamiento de la bicicleta.	Entero
idunplug_station	Número de la estación donde se desengancha la bicicleta y, por tanto, donde comienza el viaje.	Entero
ageRange	Número que indica el rango de edad del usuario que realizó el viaje. Valores posibles: 0: No se pudo determinar el rango de edad del usuario. 1: El usuario tiene entre 0 y 16 años. 2: El usuario tiene entre 17 y 18 años. 3: El usuario tiene entre 19 y 26 años. 4: El usuario tiene entre 27 y 40 años. 5: El usuario tiene entre 41 y 65 años. 6: El usuario tiene 66 años o más.	Entero
idplug_station	Número de la estación donde se ancla la bicicleta y, por tanto, donde finaliza el viaje.	Entero
unplug_hourtime	Este campo contiene a su vez un subcampo \$date que contiene fecha y hora en formato ISO 8601. Representa el momento en el cual se produce el desenganche de la bicicleta. Por motivos de anonimización de los datos, se registra exclusivamente la hora de inicio del trayecto, omitiendo la información relativa a minutos y segundos. En consecuencia, todos los desplazamientos iniciados durante una misma hora tienen asignado el mismo valor horario.	Entero
zip_code	Texto que indica el código postal del usuario que realizó el viaje.	Entero

Tabla 3.2: Descripción de los campos informados en el fichero JSON original de viajes del servicio Bicimad.

Campo	Valor
_id	{"\$_id": "5a505acd2f384319304ed614"}
user_day_code	b00665a845be18ed9f036c2d72def1ef... ...2fa2c49365cb64644821b72d560b100f
idplug_base	1
user_type	1
idunplug_base	14
travel_time	284
idunplug_station	6
ageRange	5
idplug_station	7
unplug_hourTime	{"\$date": "2018-01-01T00:00:00.000+0100"}
zip_code	28010

Tabla 3.3: Ejemplo de una entrada del fichero JSON representada en una tabla.

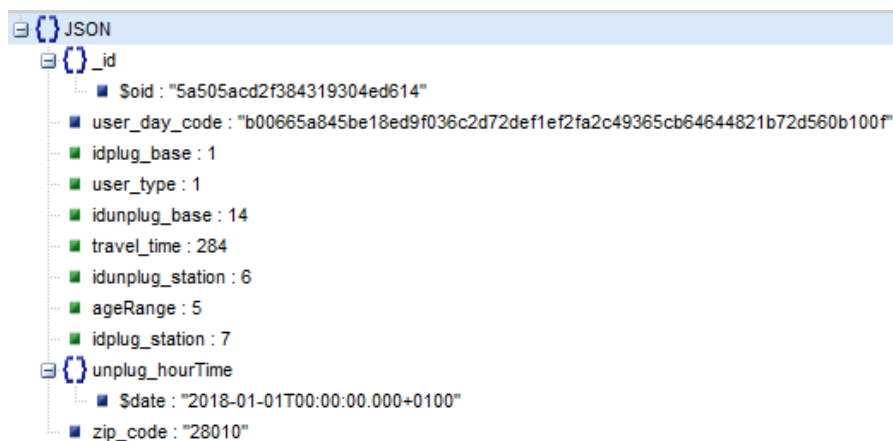


Figura 3.2: Estructura de un registro del fichero JSON de viajes. Fuente: Elaboración propia.

3.2.2. Dataset 2 - Información de estaciones de BiciMad

Este conjunto de datos representa un inventario de las estaciones del servicio BiciMAD en Madrid. Contiene coordenadas geográficas (latitud y longitud), información sobre los distritos y barrios donde se encuentran las estaciones, información de dirección específica de la estación y datos operativos como el número de plazas disponibles en cada ubicación. En la Tabla 3.4 se muestra una descripción de los campos del fichero, el tipo de dato y un ejemplo de valores de un registro. Adicionalmente, en la Imagen 3.3 se ve a modo de ejemplo algunos de los registros del fichero. El fichero inventario de las estaciones del servicio BiciMAD [EMT Madrid, 2024b] se ha obtenido del sitio web de Open Data de la Empresa Municipal de Transportes de Madrid (EMT Madrid).

Nombre del campo	Descripción	Tipo de Campo	Ejemplo
Número	Identificador único de la estación, que en algunos casos puede incluir subdivisiones. Por ejemplo, ciertas estaciones se identifican mediante un único número como por ejemplo 2, mientras que otras, debido a su tamaño y configuración, se dividen en subestaciones representadas por identificadores como 1a y 1b.	Texto	2
Gis_X	Coordenada X del sistema de información geográfica (GIS).	Decimal	440134,83
Gis_Y	Coordenada Y del sistema de información geográfica (GIS).	Decimal	4474678,23
Fecha de Alta	Fecha en que la estación fue incorporada al sistema (en formato numérico).	Fecha numérica	41813
Distrito	Distrito municipal donde se ubica la estación.	Texto	01 CENTRO
Barrio	Barrio dentro del distrito donde está la estación.	Texto	01-05 UNI- VERSIDAD
Calle	Nombre de la calle donde está ubicada la estación.	Texto	MIGUEL MOYA, CALLE, DE
Nº Finca	Número del edificio o ubicación específica en la calle.	Entero	1
Tipo de Reserva	Tipo de sistema de reserva asociado a la estación.	Texto	BiciMAD
Número de Plazas	Capacidad total de bicicletas disponibles en la estación.	Entero	24
Longitud	Coordenada de longitud en formato decimal.	Decimal	-3.7056739
Latitud	Coordenada de latitud en formato decimal.	Decimal	40.42058
Dirección	Dirección completa de la estación.	Texto	MIGUEL MOYA, CALLE, DE, 1

Tabla 3.4: Descripción de los campos del fichero inventario de las estaciones del servicio BiciMad.

bases_bicimad.csv: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
Número;Gis_X;Gis_Y;Fecha de Alta;Distrito;Barrio;Calle;Nº Finca;Tipo de Reserva;Número de Plazas;Longitud;Latitud;Dirección
001 a;440443,61;4474290,65;43803;01 CENTRO;01-06 SOL;ALCALA, CALLE, DE;2;Bicimad;30;-3.701998;40.4171107;ALCALA, CALLE, DE, 2
001 b;440480,56;4474301,74;43867;01 CENTRO;01-06 SOL;ALCALA, CALLE, DE;6;Bicimad;30;-3.7015636;40.4172132;ALCALA, CALLE, DE, 6
2;440134,83;4474678,23;41813;01 CENTRO;01-05 UNIVERSIDAD;MIGUEL MOYA, CALLE, DE;1;Bicimad;24;-3.7056739;40.42058;MIGUEL MOYA, CALLE, DE,
3;440012,98;4475760,68;41813;07 CHAMBERÍ;07-02 ARAPILES;CONDE DEL VALLE DE SUCHIL, PLAZA, DEL;2;Bicimad;18;-3.7072122;40.4303224;CONDE D
4;440396,4;4475565,36;41813;01 CENTRO;01-05 UNIVERSIDAD;MANUELA MALASAÑA, CALLE, DE;3;Bicimad;24;-3.7026739;40.4285905;MANUELA MALASAÑA,
5;440447,06;4475539,6;41813;01 CENTRO;01-04 JUSTICIA;FUENCARRAL, CALLE, DE;106;Bicimad;27;-3.7020743;40.428362;FUENCARRAL, CALLE, DE, 10
6;440754,26;4475071,08;41813;01 CENTRO;01-04 JUSTICIA;HORTALEZA, CALLE, DE;63;Bicimad;24;-3.6984093;40.4241633;HORTALEZA, CALLE, DE, 63
7;440811,97;4475187,49;41813;01 CENTRO;01-04 JUSTICIA;HORTALEZA, CALLE, DE;75;Bicimad;21;-3.6977399;40.4252161;HORTALEZA, CALLE, DE, 75
8;441020,56;4475463,16;41813;07 CHAMBERÍ;07-04 ALMAGRO;ALONSO MARTINEZ, PLAZA, DE;6;Bicimad;24;-3.6953067;40.4277143;ALONSO MARTINEZ, PL
9;439813,51;4474129,47;41813;01 CENTRO;01-01 PALACIO;SAN MIGUEL, PLAZA, DE;9;Bicimad;24;-3.7094093;40.4156134;SAN MIGUEL, PLAZA, DE, 9
10;441292,12;4475204,43;41813;01 CENTRO;01-04 JUSTICIA;MARQUES DE LA ENSENADA, CALLE, DEL;16;Bicimad;24;-3.6920816;40.4254027;MARQUES DE
11;440314,86;4475395,19;41813;01 CENTRO;01-05 UNIVERSIDAD;SAN ANDRES, CALLE, DE;20;Bicimad;24;-3.7036192;40.4270517;SAN ANDRES, CALLE, D
12;440095,53;4475560,73;41813;01 CENTRO;01-05 UNIVERSIDAD;SAN BERNARDO, CALLE, DE;87;Bicimad;24;-3.7062202;40.4285271;SAN BERNARDO, CALLE
13;439733,78;4475434,65;41813;01 CENTRO;01-05 UNIVERSIDAD;CONDE DUQUE, CALLE, DEL;22;Bicimad;24;-3.7104727;40.4273652;CONDE DUQUE, CALLE
14;439486,2;4475284,13;44045;09 MONCLOA-ARAVACA;09-02 ARGÜELLES;VENTURA RODRIGUEZ, CALLE, DE;1;Bicimad;24;-3.7133768;40.4259913;VENTURA
15;439996,77;4475313,59;43881;01 CENTRO;01-05 UNIVERSIDAD;NORTE, CALLE, DEL;10;Bicimad;21;-3.7073611;40.4262937;NORTE, CALLE, DEL, 10
```

Figura 3.3: Ejemplos de registros del fichero inventario de las estaciones del servicio BiciMad. Fuente: Elaboración propia.

3.2.3. Dataset 3 - Información meteorológica

Las condiciones meteorológicas a nivel diario se extraen recuperando datos del API de la Agencia Estatal de Meteorología de España (AEMET) publicada a través de su web [AEMET [Agencia Estatal de Meteorología de España], 2024b]. Al realizar las solicitudes correspondientes a la API [AEMET [Agencia Estatal de Meteorología de España], 2024a], se obtiene un archivo CSV que recoge información detallada sobre distintos parámetros meteorológicos por día. Entre estos se incluyen las temperaturas mínima, máxima y media, las precipitaciones, la presión atmosférica y la velocidad del viento. De entre las diversas estaciones de medición disponibles en la región metropolitana de Madrid, seleccionamos las mediciones realizadas por la estación Retiro”. De entre las estaciones de medición disponibles, esta estación, por su localización, es la que mejor representa las condiciones climáticas experimentadas por los usuarios del servicio, dadas las ubicaciones de las estaciones de BiciMad en 2018 y 2019. En la Tabla 3.5 se describen los distintos campos que ofrece el fichero CSV tras la llamada aa a API, mientras que en la Figura 3.4 se muestra un ejemplo de algunos registros de dicho fichero.

Campo	Descripción	Tipo de Datos	Ejemplo
FECHA	Fecha a la cual corresponde la medición meteorológica.	Fecha	01/01/2018
ESTACION	Nombre de la estación meteorológica.	Texto	MADRID, RETIRO
PROVINCIA	Provincia donde se ubica la estación.	Texto	MADRID
ALTITUD	Altitud de la estación en metros sobre el nivel del mar.	Entero	667
TEMP_MEDIA	Temperatura media diaria (en grados Celsius).	Decimal	7,6
PREC	Precipitación acumulada diaria (en mm).	Decimal	0
TEMP_MIN	Temperatura mínima diaria (en grados Celsius).	Decimal	3,6
HORA_TEMP_MIN	Hora de registro de la temperatura mínima. Cuando varias horas comparten la misma temperatura mínima, se indica con el texto “Varías” en lugar de una hora específica.	Hora/ Texto	08:00
TEMP_MAX	Temperatura máxima diaria (en grados Celsius).	Decimal	11,6

Continúa en la siguiente página

Campo	Descripción	Tipo de Datos	Ejemplo
HORA_TEMP_MAX	Hora de registro de la temperatura máxima. Cuando varias horas comparten la misma temperatura máxima, se indica con el texto “Varias” en lugar de una hora específica.	Hora/ Texto	13:50
DIR_VIENTO	Dirección predominante del viento (en grados).	Decimal	27
VELOCMED_VIENTO	Velocidad media del viento (en km/h).	Decimal	3,1
RACHA	Racha máxima de viento registrada (en km/h).	Decimal	10
HORA_RACHA	Hora de registro de la racha máxima. Cuando varias horas comparten la misma racha máxima, se indica con el texto “Varias” en lugar de una hora específica.	Hora/ Texto	00:10
PRES_MAX	Presión atmosférica máxima diaria (en milibares).	Decimal	954,3
HORA_PRES_MAX	Hora de registro de la presión máxima sin incluir minutos. Cuando varias horas comparten la misma presión máxima, se indica con el texto “Varias” en lugar de una hora específica.	Hora/ Texto	11
PRES_MIN	Presión atmosférica mínima diaria (en milibares).	Decimal	950,7
HORA_PRES_MIN	Hora de registro de la presión mínima sin incluir minutos. Cuando varias horas comparten la misma presión mínima, se indica con el texto “Varias” en lugar de una hora específica.	Hora/ Texto	16

Tabla 3.5: Descripción de los campos del dataset meteorológico de la estación Retiro.

Temperaturas Retiro.csv: Bloc de notas

Archivo Edición Formato Ver Ayuda

```
FECHA;ESTACION;PROVINCIA;ALTITUD;TEMP_MEDIA;PREC;TEMP_MIN;HORA_TEMP_MIN;TEMP_MAX;HORA_TEMP_MAX;DIR_VIENTO;VELOCMED_VIENTO;RACHA;HORA_RACHA;SOL;PRES_MAX;HORA_PRES_MAX;PRES_MIN;HORA_PRES_MIN
01/01/2018;MADRID; RETIRO;MADRID;667;7,6;0;3,6;08:00;11,6;13:50;27;3,1;10:00;0;954,3;11;950,7;16
02/01/2018;MADRID; RETIRO;MADRID;667;10,4;0;7,5;04:10;13,4;13:30;29;2,5;11,4;02:50;0;954,3;Varias;952,2;Varias
03/01/2018;MADRID; RETIRO;MADRID;667;10,6;0;7,6;02:15;13,5;14:10;22;1,7;8,6;03:40;0;955,4;11;952,3;24
04/01/2018;MADRID; RETIRO;MADRID;667;10,8;0;7,8;Varias;13,7;13:20;99;2,2;7,8;Varias;0;953,3;00;943,7;24
05/01/2018;MADRID; RETIRO;MADRID;667;8,6;14,1;5,8;23:59;11,5;13:50;22;3,1;11,7;07:50;0;943,7;00;926,1;24
06/01/2018;MADRID; RETIRO;MADRID;667;4,20,6;2;23:59;6,1;00:00;3;2,5;10,8;09:50;0;926,1;00;920;14
07/01/2018;MADRID; RETIRO;MADRID;667;3,3;3,1;1,6;23:59;5,13;40;12;2,8;7,5;11:50;0;929,5;23;919,5;05
08/01/2018;MADRID; RETIRO;MADRID;667;3,8;0;0,7;23:50;7;13:50;4;1,4;7,2;05:10;0;937,8;24;929,3;00
09/01/2018;MADRID; RETIRO;MADRID;667;1,6;2,3;-1,4;05:20;4,7;15:15;19;1,4;4,7;16:50;0;941,6;10;937,2;01
10/01/2018;MADRID; RETIRO;MADRID;667;5,1;0,8;3,2;00:30;7;16:00;23;2,2;9,4;15:20;0;943;24;939,5;05
11/01/2018;MADRID; RETIRO;MADRID;667;5,8;0;2,9;08:10;8,8;14:40;30;3,3;12,8;11:40;0;946,6;24;942,6;05
12/01/2018;MADRID; RETIRO;MADRID;667;3,8;0;-0,3;07:40;7,9;16:10;28;1,7;6,4;00:20;0;947,8;10;944,3;24
13/01/2018;MADRID; RETIRO;MADRID;667;4,6;2,4;04:20;5,6;12:10;26;1,1;7,2;21:20;0;944,5;00;939;17
14/01/2018;MADRID; RETIRO;MADRID;667;4,6;0;2,5;02:20;6,7;14:50;3;1,7;6,1;20:50;0;944,4;24;939,8;06
15/01/2018;MADRID; RETIRO;MADRID;667;4,2;0;07:30;8,5;15:00;22;1,1;6,4;15:40;0;949,9;24;944,4;00
16/01/2018;MADRID; RETIRO;MADRID;667;7,2;0;2;07:30;12,5;16:30;26;1,1;6,4;22:10;0;951,2;Varias;947,8;18
17/01/2018;MADRID; RETIRO;MADRID;667;9;0;4,8;08:30;13,2;14:40;30;2,2;8,6;01:00;0;948,8;Varias;945,9;06
18/01/2018;MADRID; RETIRO;MADRID;667;7,8;0;2,7;08:00;12,8;16:10;3;1,4;5,3;08:50;0;948,4;00;943,1;24
```

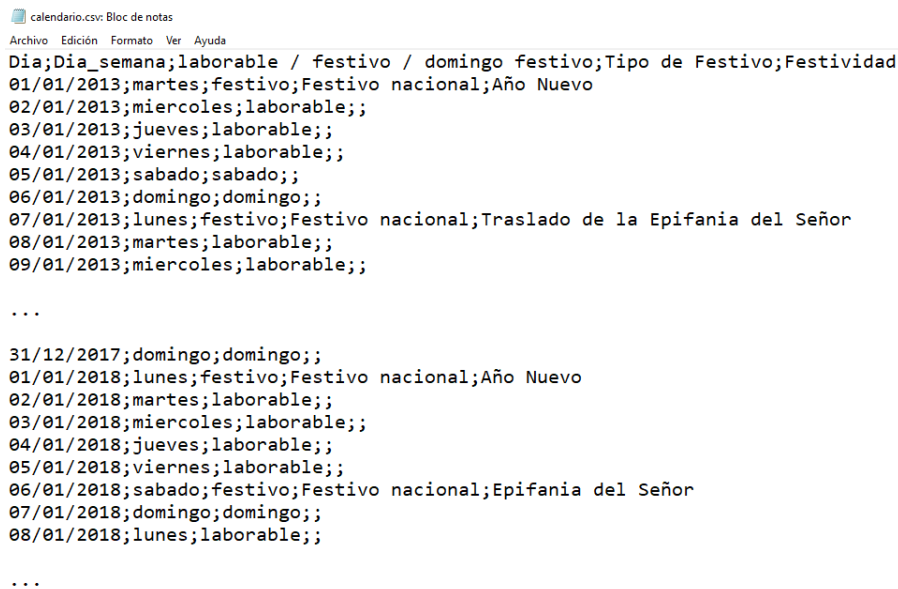
Figura 3.4: Subconjunto de registros a modo de ejemplo del fichero de valores meteorológico registrados por la estación de medición de Retiro. Fuente: Elaboración propia.

3.2.4. Dataset 4 - Calendario de días festivos

Para poder clasificar las distintas fechas como laborables o festivos/ fin de semana se utiliza el fichero *calendario.csv*. Dicho fichero es el calendario laboral de la ciudad de Madrid, extraído del portal de datos abiertos del Ayuntamiento de Madrid [[Ayuntamiento de Madrid, Portal de Datos Abiertos, 2024](#)] el cual incluye los festivos municipales, autonómicos y nacionales que afectan a la ciudad. La Tabla 3.6 muestra la estructura de datos del fichero, mientras que en la Figura 3.5 se presentan algunos registros a modo de ejemplo.

Nombre del campo	Descripción	Tipo	Ejemplo
Dia	Fecha en formato DD/MM/AAAA	Fecha	01/01/2018
Dia_semana	Nombre del día de la semana	Texto	lunes
laborable /	Indica si el día es laborable, festivo o	Texto	festivo
festivo /	domingo festivo.		
Tipo de Festivo	Clasificación del festivo.	Texto	Festivo nacional
Festividad	Nombre o descripción del festivo	Texto	Año Nuevo

Tabla 3.6: Descripción de los campos del fichero calendario.csv.



```

calendario.csv: Bloc de notas
Archivo Edición Formato Ver Ayuda
Dia;Dia_semana;laborable / festivo / domingo festivo;Tipo de Festivo;Festividad
01/01/2013;martes;festivo;Festivo nacional;Año Nuevo
02/01/2013;miercoles;laborable;;
03/01/2013;jueves;laborable;;
04/01/2013;viernes;laborable;;
05/01/2013;sabado;sabado;;
06/01/2013;domingo;domingo;;
07/01/2013;lunes;festivo;Festivo nacional;Traslado de la Epifania del Señor
08/01/2013;martes;laborable;;
09/01/2013;miercoles;laborable;;

...

31/12/2017;domingo;domingo;;
01/01/2018;lunes;festivo;Festivo nacional;Año Nuevo
02/01/2018;martes;laborable;;
03/01/2018;miercoles;laborable;;
04/01/2018;jueves;laborable;;
05/01/2018;viernes;laborable;;
06/01/2018;sabado;festivo;Festivo nacional;Epifania del Señor
07/01/2018;domingo;domingo;;
08/01/2018;lunes;laborable;;

...

```

Figura 3.5: Subconjunto de registros del fichero calendario.csv a modo de ejemplo

Fuente: Elaboración propia.

3.2.5. Dataset 5 - Estado en tiempo real de las estaciones de BiciMad

El último dataset utilizado en esta tesis se genera a partir de consultas realizadas a la API de BiciMad [EMT Madrid, 2024]. Cada consulta produce un archivo JSON que facilita la información en tiempo real sobre el estado de todas las estaciones. La Tabla 3.7 ofrece una descripción detallada de los campos del archivo JSON, mientras que la Figura 3.6 muestra un ejemplo de un registro representativo del estado en tiempo real de una estación en un instante determinado. Este conjunto de datos será clave en el Capítulo 5. para desarrollar una prueba de concepto de una aplicación práctica del modelo desarrollado en esta tesis.

Campo	Descripción	Tipo de Campo	Ejemplo
id	Código de la estación.	Entero	1

Continúa en la siguiente página

Campo	Descripción	Tipo de Campo	Ejemplo
name	Nombre descriptivo de la estación. Puede incluir información adicional como subdivisiones o el nombre de la ubicación.	Texto	1a - Puerta del Sol A
light	Grado de Ocupación (0=baja, 1=media, 2=alta, 3=inactiva).	Entero	2
number	Denominación lógica de la estación, que puede incluir subdivisiones si la estación es grande.	Entero	1a
address	Dirección física de la estación.	Texto	Puerta del Sol nº 1
activate	Estación activa (0=No activa, 1=activa).	Entero	1
no_available	Disponibilidad de la Estación (0=disponible, 1=no disponible).	Entero	0
total_bases	Número de bases (punto de anclaje para bicicletas) de la estación.	Entero	25
dock_bikes	Número de bicicletas actualmente ancladas en la estación.	Entero	15
free_bases	Número de bases libres en la estación.	Entero	10
reservations_count	Número de reservas activas en la estación.	Entero	2
geometry	Objeto que contiene la información geográfica de la estación. Está compuesto a su vez por un subcampo <i>type</i> que indica el tipo de geometría (en este caso se trata de un punto geográfico) y otro subcampo denominado <i>coordinates</i> que contiene las coordenadas geográficas de la estación en formato [longitud, latitud] siguiendo el sistema WGS84.	Objeto	type: "Point" coordinates: [-3.7018341, 40.4172137]

Tabla 3.7: Descripción de los campos del archivo JSON proporcionado por la API de estado en tiempo real de las estaciones de BiciMad.

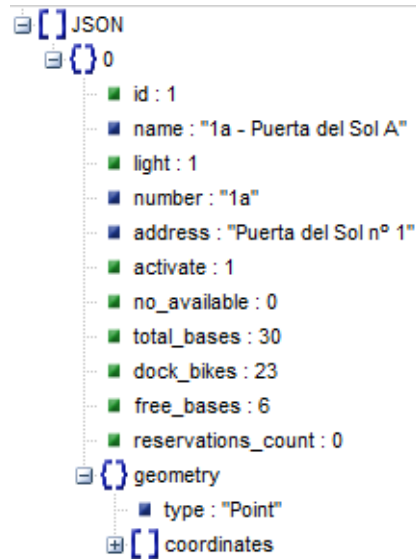


Figura 3.6: Ejemplo de un registro del fichero JSON con información de estaciones de BiciMad en tiempo real Fuente: Elaboración propia.

3.3. Extracción y transformación de los datos

El proceso de extracción y transformación de los datos organiza y prepara la información necesaria para poder realizar el análisis de la misma y emplear los datos en el modelo que se explicará en el Capítulo 4. Para este estudio se diseñó un flujo de trabajo riguroso, que permitió obtener los datos de la fuente original, enriquecerlos con información de fuentes adicionales, sanitizarlos, prepararlos y formatearlos para su posterior procesamiento. La Figura 3.7 ofrece una representación visual de este proceso, que será desglosado y explicado en las secciones siguientes.

Finalmente, cabe mencionar que se optó por describir los procesos de extracción y transformación como un flujo continuo para facilitar su explicación. Sin embargo, es importante señalar que el análisis exploratorio de datos (EDA), realizado entre ambas etapas y presentado en el capítulo siguiente, permitió identificar la necesidad de incorporar algunas transformaciones adicionales al proceso.

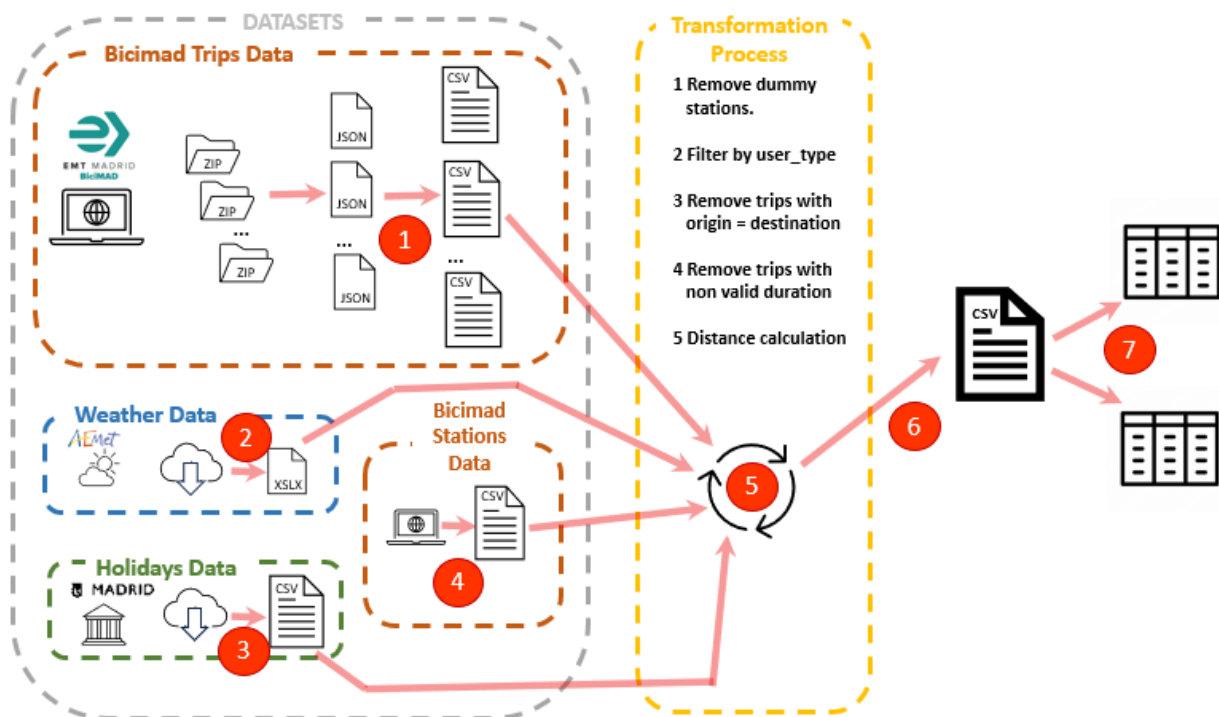


Figura 3.7: Diagrama del flujo de trabajo para la extracción y transformación de los datos procedentes de las fuentes originales. Fuente: Elaboración propia.

3.3.1. Proceso de extracción de los datos de viajes

En el paso inicial, se realizó la extracción de los datos de viajes de BiciMad desde el portal oficial de datos abiertos de EMT Madrid [datos abiertos](#). Este proceso se llevó a cabo utilizando el paquete de Python *BeautifulSoup* para ejecutar web scraping, obteniendo los archivos ZIP originales. Una vez descargados, los archivos ZIP se descomprimieron para acceder a los datos en formato JSON, que fueron posteriormente procesados y convertidos en archivos CSV. El procedimiento es explicado en el Algoritmo 1.

3.3.2. Proceso de extracción de los datos del resto de fuentes

La extracción de datos del inventario genérico de estaciones de BiciMad (Dataset 2) se realizó de forma directa desde la [URL](#), sin necesidad de automatización. El acceso a la información es inmediato mediante la descarga directa desde la fuente.

En el caso del dataset 3 con información meteorológica las peticiones a la API se realizan directamente a través de la [web](#) mediante un proceso de Rest API, obteniendo como resultado la posibilidad de descargar manualmente la información en un fichero CSV.

El proceso para extraer el Dataset 4, relacionado con el calendario de días festivos, es análogo al empleado para el Dataset 2. Los datos se obtienen accediendo a la [URL](#) y están disponibles para descarga directa, sin necesidad de automatización.

Finalmente, la extracción de los datos de ocupación e información general de las estaciones de BiciMad en tiempo real se llevó a cabo mediante un programa desarrollado en Python, que realiza solicitudes a la [API](#) correspondiente.

Los pasos principales realizados en este programa son:

Algoritmo 1: Proceso de extracción de datos de viajes de BiciMad**Input** : HTML source URL:`https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)`**Output:** CSV files containing processed data**foreach** $i \in \text{lines of source HTML}$ **do** **if** $i.\text{labeltype} = 'a'$ **and** $i.\text{href} = \text{True}$ **then** **if** `"getattachment"` **and** (`usage` **or** `movements`) **in** $i.\text{hrefcontent}$ **then** `listlinks.add(i)` // Add URL to list of movement data files**foreach** $\text{movementdatalink} \in \text{listlinks}$ **do** $\text{zipfile} \leftarrow \text{request.geturl}(\text{movementdatalink});$ $\text{jsonfile} \leftarrow \text{unzip}(\text{zipfile});$ $\text{dataframe} \leftarrow \text{read_json}(\text{jsonfile});$ $\text{csvfile} \leftarrow \text{ProcessDataFrameToCSV}(\text{dataframe});$ `WriteCSV(csvfile);`

- Autenticación inicial con la API mediante un token de acceso (`accessToken`).

```

1      # Authentication
2      response = requests.request("GET", URL_LOGIN, headers=headers)
3      response_json = json.loads(response.text)
4      accessToken = response_json["data"][0]["accessToken"]
5      headers_token = {"accessToken": accessToken}
6
7      URL_STATIONS =
      ↪ "https://openapi.emtmadrid.es/v1/transport/bicimad/stations/"

```

- Realización de múltiples solicitudes a la API en intervalos regulares 5 minutos para obtener el estado en tiempo real de las estaciones. Dado que nuestro modelo opera en base horaria, posteriormente se calcula el promedio de bicicletas disponibles y de anclajes vacíos por hora y por estación. La Figura 3.8 muestra un ejemplo de la salida ofrecida por el programa al realizar tres peticiones cada 5 minutos.

3. Análisis exploratorio de los datos (EDA)

```
Number of measurement: = 1
{'Server': 'unicorn/20.0.4', 'Date': 'Fri, 02 Jun 2023 15:07:04 GMT', 'Content-Type': 'application/json', 'Content-Length': '129421',
'Access-Control-Allow-Origin': '*', 'X-INSTANA-L': '1', 'traceparent': '00-0000000000000000c1c4b2365d0dc45b-c1c4b2365d0dc45b-01',
'tracestate': 'in=c1c4b2365d0dc45b;c1c4b2365d0dc45b', 'X-INSTANA-T': 'c1c4b2365d0dc45b', 'X-INSTANA-S': 'c1c4b2365d0dc45b',
'Server-Timing': 'intid;desc=c1c4b2365d0dc45b', 'Set-Cookie': 'SERVERIDP=4990ba58bc04669309648557c6058ecfb5466768; path=/, 'Cache-control': 'private'}
date = 02/06/2023
time = 17:07:03
waiting 5 minutes ...
Number of measurement: = 2
{'Server': 'unicorn/20.0.4', 'Date': 'Fri, 02 Jun 2023 15:12:04 GMT', 'Content-Type': 'application/json', 'Content-Length': '129420',
'Access-Control-Allow-Origin': '*', 'X-INSTANA-L': '1', 'traceparent': '00-0000000000000000f3ed727caa5c888a-f3ed727caa5c888a-01',
'tracestate': 'in=f3ed727caa5c888a;f3ed727caa5c888a', 'X-INSTANA-T': 'f3ed727caa5c888a', 'X-INSTANA-S': 'f3ed727caa5c888a',
'Server-Timing': 'intid;desc=f3ed727caa5c888a', 'Set-Cookie': 'SERVERIDP=4990ba58bc04669309648557c6058ecfb5466768; path=/, 'Cache-control': 'private'}
date = 02/06/2023
time = 17:12:04
waiting 5 minutes ...
Number of measurement: = 3
{'Server': 'unicorn/20.0.4', 'Date': 'Fri, 02 Jun 2023 15:17:06 GMT', 'Content-Type': 'application/json', 'Content-Length': '129415',
'Access-Control-Allow-Origin': '*', 'X-INSTANA-L': '1', 'traceparent': '00-0000000000000000f9c7ca95611b210b-f9c7ca95611b210b-01',
'tracestate': 'in=f9c7ca95611b210b;f9c7ca95611b210b', 'X-INSTANA-T': 'f9c7ca95611b210b', 'X-INSTANA-S': 'f9c7ca95611b210b',
'Server-Timing': 'intid;desc=f9c7ca95611b210b', 'Set-Cookie': 'SERVERIDP=8c1468057c50abe38d26ad988f6b7794baffae99; path=/, 'Cache-control': 'private'}
date = 02/06/2023
time = 17:17:05
waiting 5 minutes ...
Number of measurement: = 4
...
```

Figura 3.8: Ejemplos de salidas del programa que realiza múltiples solicitudes a la API en intervalos de 5 minutos. Fuente: Elaboración propia.

- Procesamiento de las respuestas JSON para extraer la información relevante de cada estación, lo cual se lleva a cabo utilizando la función:

```
1 def extract_bicimad_station(data):
2     return [
3         data["id"],
4         data["number"],
5         data["light"],
6         data["activate"],
7         data["no_available"],
8         data["total_bases"],
9         data["dock_bikes"],
10        data["free_bases"],
11        data["reservations_count"],
12        data["name"]
13    ]
```

- Almacenamiento de los datos extraídos en un archivo CSV y en una base de datos MongoDB. La idea de almacenarlo en un fichero CSV es para su posterior procesos mientras que la de almacenarlo en MongoDB era la de tener un respaldo fiel a la extracción e ir contruyendo un histórico de información. La Figura 3.9 muestra un ejemplo de registro guardado en MongoDB mientras que la Figura 3.10 muestra una serie de registros a modo de ejemplo guardados en el fichero CSV.

```
{'_id': ObjectId('622f7349d1fdb79bf7f9b356'), 'meditation_id': 1, 'date': '14/03/2022',
'time': '17:54:33', 'info': {'id': 1, 'name': '1a - Puerta del Sol A', 'light': 1,
'number': '1a', 'address': 'Puerta del Sol nº 1', 'activate': 1, 'no_available': 0,
'total_bases': 30, 'dock_bikes': 22, 'free_bases': 6, 'reservations_count': 0,
'geometry': {'type': 'Point', 'coordinates': [-3.7018341, 40.4172137]}}
```

Figura 3.9: Ejemplo de registro guardado en MongoDB. Fuente: Elaboración propia.

```

1;29/03/2022;15:54:43;1;1a;2;1;0;30;16;14;0;1a - Puerta del Sol A
1;29/03/2022;15:54:43;2;1b;2;1;0;30;10;20;0;1b - Puerta del Sol B
1;29/03/2022;15:54:43;3;2;2;1;0;24;9;15;0;2 - Miguel Moya
1;29/03/2022;15:54:43;4;3;2;1;0;18;7;11;0;3 - Plaza Conde Suchil
1;29/03/2022;15:54:43;5;4;2;1;0;24;6;13;0;4 - Malasaña
1;29/03/2022;15:54:43;6;5;2;1;0;27;11;15;0;5 - Fuencarral
1;29/03/2022;15:54:43;7;6;1;1;0;24;17;7;0;6 - Colegio Arquitectos
1;29/03/2022;15:54:43;8;7;2;1;0;21;12;9;0;7 - Hortaleza
...

```

Figura 3.10: Ejemplo de registros de información en tiempo real de las estaciones guardado en fichero CSV. Fuente: Elaboración propia.

- Revisión continua de un archivo de control para determinar si el programa debe continuar ejecutándose, funcionando como un mecanismo de control para detener el bucle infinito de *polling* cuando sea necesario.

El algoritmo del programa desarrollado se presenta en el bloque de Algoritmo 2.

Algoritmo 2: Proceso de extracción de Datos de las Estaciones de BiciMad en tiempo real

Data: Control file `filecontinue.txt`, Authentication URL `URL_LOGIN`, Stations URL `URL_STATIONS`

Result: Data stored in a CSV file and MongoDB

Authentication: Perform a GET request to `URL_LOGIN` to obtain the `accessToken`;

Set up `headers_token` with the `accessToken`;

Open `filecontinue.txt` and read the initial execution state (`continuequest`);

while `continuequest = "Yes"` **do**

 Increment the measurement counter `num_m`;

 Perform a GET request to `URL_STATIONS` and process the JSON response (`parsed`);

 Get the current date and time (`now`);

 Open the CSV file in append mode;

foreach `station in parsed` **do**

 Extract key data using the function `extract_bicimad_station`;

 Write the data to the CSV file and store it in MongoDB;

 Close the CSV file;

 Wait 5 minutes before the next iteration;

 Read the state of `filecontinue.txt` to determine if the process should continue;

3.3.3. Procesos de transformación

La tabla 3.8 muestra de forma esquematizada las transformaciones que se han aplicado a los datos extraídos de las fuentes originales. Dichas transformaciones corresponden con el paso 5 de la Figura 3.7.

Id	Tipo de Transformación	Campos Afectados/Involucrados	Descripción / Comentarios
1	Eliminar identificadores de estaciones de prueba (<i>dummy</i>)	idunplug_station e idplug_station	Según la información proporcionada por EMT en el portal de datos abiertos, durante la expansión de la cobertura de estaciones de BiciMad se llevaron a cabo muchas pruebas. Algunas de ellas se incluyeron en el conjunto de datos publicado. Por ello, fue necesario eliminar los viajes que incluían estaciones de laboratorio (denotadas por los números 1000, 1001, del 2000 al 2017, y 3000).
2	Filtrado por el campo user_type	user_type	El tipo de usuario que realizó la transacción está indicado por el valor numérico en la columna user_type. Filtramos esta variable por el valor 1, que corresponde a Usuario Anual (titular de un abono anual), de acuerdo con el alcance de nuestro estudio.
3	Eliminar viajes con estación de origen igual a la de destino	idunplug_station e idplug_station	Durante el análisis exploratorio de datos del conjunto de viajes de Bicimad, se encontraron viajes con estaciones de origen y destino iguales. Se identificaron dos situaciones para este tipo de casos: La primera tiene que ver con problemas al desanclar las bicicletas. Cuando el usuario intenta tomar una bicicleta que no se libera del punto de anclaje de la estación o tiene algún problema, termina dejándola nuevamente en el mismo lugar. Estos viajes suelen durar solo unos segundos, por lo que son fáciles de identificar. La segunda situación involucra estaciones situadas en destinos turísticos, como parques. Normalmente, el usuario toma otra forma de transporte hasta la estación, recoge una bicicleta, pasea por el área de interés y luego regresa al punto de partida para recoger su modo de transporte original.

Id	Tipo de Transformación	Campos Afectados/Involucrados	Descripción / Comentarios
4	Eliminar viajes con una duración no válida	travel_time	Se considera que los viajes con una duración inferior a 1 minuto o superior a 3 horas deben ser filtrados y descartados. Generalmente, las duraciones inferiores a 1 minuto suelen deberse a errores al desanclar las bicicletas, mientras que las superiores a 3 horas pueden estar relacionadas con problemas del sistema que no registraron correctamente el anclaje final, o casos de vandalismo en los que la bicicleta fue abandonada sin ser anclada.
5	Cálculo de la distancia del viaje	idunplug_station, idplug_station, longitud y latitud de las estaciones de origen y destino	El cálculo de la distancia recorrida en un viaje se realiza utilizando las coordenadas de longitud y latitud de las estaciones de origen y destino. Dado que existen diversos métodos para realizar este cálculo, en esta investigación se ha optado por emplear una combinación de dos enfoques posibles. Por ello, resulta necesario proporcionar una explicación detallada de esta transformación, la cual se presenta en el subcapítulo 3.3.3.1 .

Tabla 3.8: Listado y descripción breve de las transformaciones realizadas a los datos originales

3.3.3.1. Método de cálculo de distancias en los viajes de BiciMad

Existen varios métodos para calcular la distancia entre dos puntos geográficos definidos por sus respectivas latitudes y longitudes. Los métodos de cálculo de distancia más comúnmente utilizados se describen en la Tabla [3.9](#).

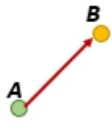
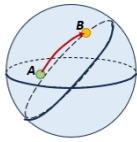
Nombre	Descripción y Fórmula
Euclidiana 	Es la distancia en línea recta entre dos coordenadas en el espacio euclidiano. Se calcula utilizando el teorema de Pitágoras. Para un espacio euclidiano de n dimensiones: $d_E(p, q) = \sqrt{\sum_{i=1}^N (p_i - q_i)^2}$ Esta fórmula puede particularizarse asumiendo un espacio bidimensional (latitud, longitud): $d_E(p, q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ donde (x_1, y_1) son las coordenadas del punto p y (x_2, y_2) las del punto q, respectivamente.
Manhattan 	Es una medida de la distancia entre dos puntos en una cuadrícula, como un plano urbano, y también se conoce como la norma de taxista. $d_M(p, q) = p_1 - q_1 + p_2 - q_2 $
Haversine 	Es la distancia entre dos puntos en la superficie de una esfera. Si asumimos que la Tierra es una esfera, obtenemos la siguiente fórmula: $d_H(p, q) = 2 * R * \arcsin \sqrt{\sin^2 \left(\frac{lat1 - lat2}{2} \right) + \cos(lat1) * \cos(lat2) * \sin^2 \left(\frac{long1 - long2}{2} \right)}$ donde R, en este caso, es igual a 6371 km, que es el radio de la Tierra.

Tabla 3.9: Métodos para calcular la distancia entre dos puntos geográficos

Varios estudios académicos han analizado la efectividad de distintos métodos para calcular distancias geográficas, como se menciona en [Cahyono et al., 2022]. Por su parte, una revisión de la literatura muestra que numerosos trabajos sobre sistemas de bicicletas compartidas tienden a preferir la distancia Haversine (por ejemplo, [Li et al., 2019b]), incluso tras evaluar otras opciones. De forma comparable, investigaciones en otros modos de transporte, como los taxis [Poongodi et al., 2022], también confirman la elección del método Haversine para medir distancias.

Para nuestro estudio, era importante distinguir la dirección del viaje entre dos puntos. Por ello, resultaba necesario que la distancia entre A y B fuese ligeramente distinta de la distancia entre B y A. Asimismo, esta diferencia debía ser lo suficientemente mínima como para no influir significativamente en los resultados del modelo. Por esta razón, optamos por utilizar una combinación de las distancias Haversine y Manhattan.

Esta nueva forma de calcular la distancia entre dos puntos, que representan cada uno la localización de una estación de Bicima, se divide en dos tramos para capturar de forma más detallada el recorrido. En el primer tramo, se calcula la distancia en línea curva desde el punto de origen hasta un punto intermedio que

comparte la misma latitud que el punto de origen, pero con la longitud del punto de destino. A continuación, se calcula el segundo tramo, que conecta el punto intermedio con el punto de destino. Ambos tramos se suman para obtener la distancia final.

Este enfoque permite emular un trayecto más realista, como el que se realizaría sobre una cuadrícula urbana, parte que corresponde a la distancia Manhattan, pero utilizando la distancia Haversine como base para ambos tramos, en lugar de calcular la distancia lineal.

Como se indicó anteriormente, este enfoque permite obtener distancias distintas para un trayecto del punto A al punto B en comparación con el trayecto inverso, del punto B al punto A, debido a la curvatura de la Tierra. La Figura 3.11 ilustra este método de cálculo de distancias, mientras que el Algoritmo 3 presenta los cálculos realizados en formato de pseudocódigo.



Figura 3.11: Representación gráfica de cálculo de la distancia basado en una combinación de las distancias Haversine y Mahattan. Fuente: Elaboración propia.

Algoritmo 3: Cálculo de la distancia entre dos estaciones utilizando combinación de Haversine y Manhattan

Data: Latitud y longitud del punto de origen ($lat1, lon1$) y del destino ($lat2, lon2$).

Result: Distancia total entre el punto de origen y el destino.

Paso 1: Calcular la distancia del *tramo 1*, desde ($lat1, lon1$) hasta un punto intermedio ($lat1, lon2$) usando la fórmula de Haversine:

$$d_1 = R \cdot 2 \cdot \arcsin \sqrt{\sin^2 \left(\frac{lat1 - lat1}{2} \right) + \cos(lat1) \cdot \cos(lat1) \cdot \sin^2 \left(\frac{lon2 - lon1}{2} \right)}$$

Paso 2: Calcular la distancia del *tramo 2*, desde ($lat1, lon2$) hasta el punto final ($lat2, lon2$) usando la misma fórmula:

$$d_2 = R \cdot 2 \cdot \arcsin \sqrt{\sin^2 \left(\frac{lat2 - lat1}{2} \right) + \cos(lat1) \cdot \cos(lat2) \cdot \sin^2 \left(\frac{lon2 - lon2}{2} \right)}$$

Paso 3: Sumar ambas distancias para obtener la distancia total:

$$\text{Distancia total} = d_1 + d_2$$

Nota: R es el radio de la Tierra (6373 km) y las distancias se convierten a metros multiplicándolas por 1000.

3.4. Análisis exploratorio de variables relevantes

El análisis exploratorio de datos (*EDA*) presentado en este capítulo se centra exclusivamente en aquellas variables consideradas de mayor relevancia.

Hay que señalar que este análisis exploratorio corresponde a los datos antes de aplicar las diversas transformaciones descritas en este capítulo.

La Figura 3.12 muestra el histograma de número de usuarios por rango de edad (variable `age_range` dividida en 6 rangos según se detalla en la Tabla 3.2).

El grupo de edad que más viajeros presenta (aproximadamente 1.4 millones es el 0, el cual corresponde a aquellos usuarios cuyo rango de edad no está identificado. El análisis también muestra que los usuarios del grupo de edad entre 27 y 40 años (grupo 4) son los siguientes más numerosos seguidos de los de entre 41 y 65 años (grupo 5). Adicionalmente hay que tener en cuenta que los intervalos de edad definidos en los datos iniciales no abarcan igual número de años.

En el modelo que se desarrolla en el Capítulo 4 no incorpora esta variable, debido a que el enfoque del mismo es representar dinámicas poblacionales en vez de individuales (ver Figura 1.8). No obstante aporta una idea de la heterogeneidad de los datos iniciales.

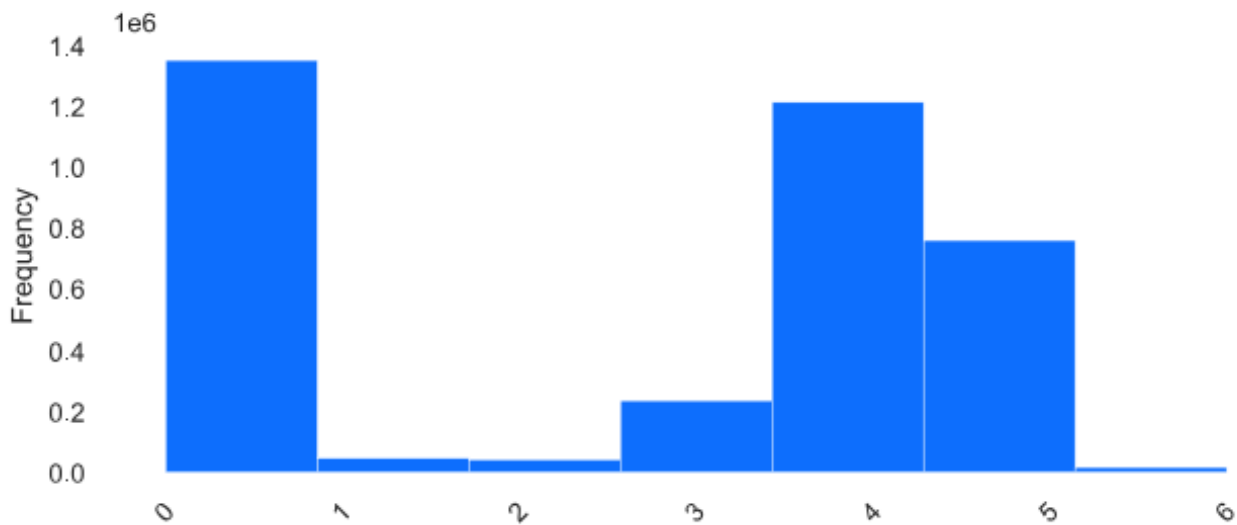


Figura 3.12: Histograma de número de usuarios por rango de edad.

La Tabla 3.10 muestra la frecuencia de cada uno de los valores que puede tener la variable `age_range` ordenados por frecuencia de mayor a menor.

Valor <code>age_range</code>	Conteo	Frecuencia (%)
0	1.355.749	36,9 %
4	1.217,575	33,1 %
5	763.320	20,8 %
3	236.331	6,4 %
1	46.335	1,3 %
2	41.848	1,1 %
6	16.928	0,5 %

Tabla 3.10: Distribución de valores, conteos y frecuencias relativas de la variable `age_range`.

Seguidamente se analiza la variable `user_type`. Hay que tener en cuenta que en esta tesis el alcance queda limitado a los registros con valor igual a 1 (usuario anual como se refleja en la Tabla 3.2) en este campo.

El objetivo de analizar esta variable es mostrar que los valores que no se tienen en cuenta, además de representar tipologías de usuarios cuya inclusión en el modelo no aplica, son poco significativos en lo que al número de registros se refiere como se aprecia en la Tabla X.

Valor <code>user_type</code>	Conteo	Frecuencia (%)
1	3.327.888	90,5 %
3	288.111	7,8 %
2	59.990	1,6 %
0	2.097	0,1 %

Tabla 3.11: Distribución de valores, conteos y frecuencias relativas de la variable `use_type`.

Finalmente, la Figura 3.13 presenta el histograma de distribución del número de viajes por hora. Tal como era previsible, la gráfica refleja la temporalidad característica de estos servicios, evidenciando una dis-

minución en la actividad durante las horas de la madrugada y un aumento pronunciado en los viajes durante los horarios comúnmente considerados de mayor demanda.

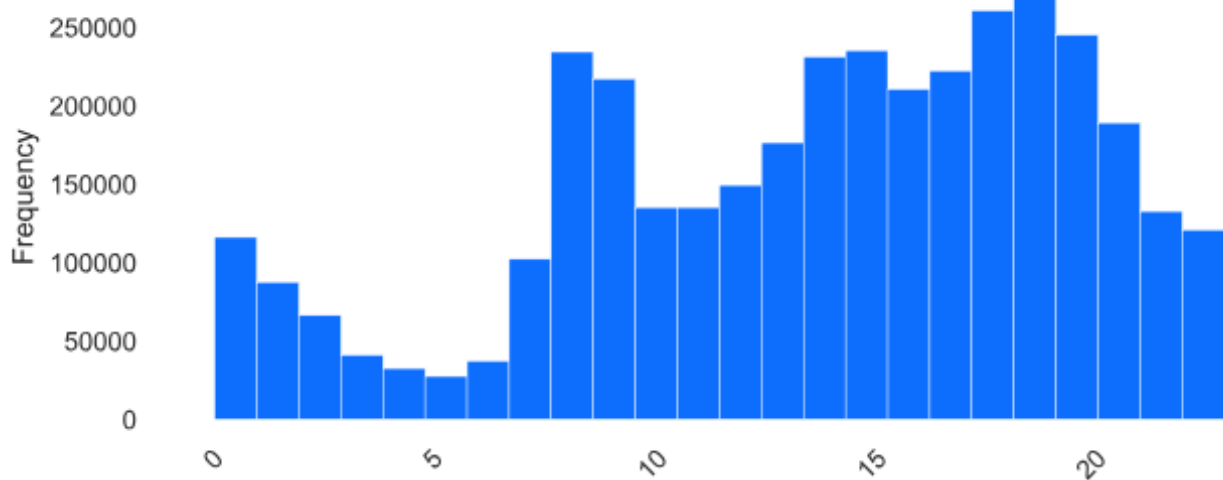


Figura 3.13: Histograma de número de viajes por hora.

3.5. Resultados y conclusiones

En este capítulo inicialmente se define el alcance de los datos empleados. Seguidamente se describen las fuentes y formatos de los datos iniciales y se describen los procesos de extracción y transformación empleados. Es destacable la justificación realizada para no emplear ciertas variables en el modelo, al tratarse de un enfoque poblacional. No obstante se analizan dichas variables para aportar una visión de lo heterogéneos que son los datos. Tras realizar esas transformaciones se cuenta finalmente con 3.249.624 registros.

Con los análisis y descripciones incluidas en este capítulo se responde a la pregunta de investigación RQ3 y sus correspondientes subpreguntas descritas en el Capítulo 1.5.2, cumpliendo también con los objetivos establecidos OB3.1 y OB3.2 detallado en el Capítulo 1.5.3.

“Les questions les plus importantes de la vie ne sont en effet, pour la plupart, que des problèmes de probabilité.”

Las preguntas más importantes de la vida, de hecho, no son en su mayoría más que problemas de probabilidad.

— Pierre-Simon Laplace *Ensayo filosófico sobre las probabilidades (Essai philosophique sur les probabilités)*, 1814

4

Enfoque de ML probabilístico para el modelado de la demanda en BiciMad

Índice del capítulo

4.1. Inferencia de las distribuciones de probabilidad	77
4.2. Modelado de la distancia	78
4.3. Modelado del número de viajes	88
4.4. Creación de un simulador probabilístico	92
4.5. Descripción de alternativas para el ajuste de distribuciones con enfoque bayesiano	94
4.6. Resultados y conclusiones	97

En este capítulo se desarrolla un modelo de aprendizaje automático probabilístico para representar y simular la demanda del servicio de bicicletas compartidas BiciMad. La demanda se modela mediante el muestreo conjunto de la distancia recorrida y el número de viajes realizados en esa distancia, utilizando dos distribuciones de probabilidad. Las distancias recorridas en los viajes se modelan con una distribución continua Gamma, fundamentada en el principio de máxima entropía, que, como demostramos, captura con precisión la distribución empírica en todos los rangos de distancias.

Además, analizamos el número de viajes por hora utilizando distribuciones discretas binomiales negativas (NegBin), considerando factores como días laborales, festivos y condiciones climáticas. Esta distribución se deriva de la mezcla de distribuciones de Poisson con tasas variables, siendo la más parsimoniosa para modelar eventos de conteo.

Integrando ambas distribuciones y añadiendo información sobre la popularidad de las estaciones, desarrollamos un simulador probabilístico que permite predecir escenarios hipotéticos y realizar análisis contrafactuales, facilitando la toma de decisiones en la planificación del servicio.

Este capítulo aborda las siguientes preguntas: *¿Cómo se puede modelar la demanda en el servicio de BiciMad? ¿Qué tipos de algoritmos de aprendizaje automático pueden usarse para cuantificar e incluir la incerti-*

dumbre en los modelos? ¿Cómo justificar la validez del modelo? ¿Se puede crear un modelo de predicción de la demanda para BiciMad que sea explicativo y contemple la incertidumbre?

4.1. Inferencia de las distribuciones de probabilidad

Para predecir la demanda del servicio, ajustamos los datos empíricos relacionados con la distancia recorrida y el número de viajes realizados en BiciMad utilizando la técnica de Estimación de Máxima Verosimilitud (*MLE*, por sus siglas en inglés de “*Maximum Likelihood Estimation*”). Este método, ampliamente empleado en estadística clásica y en aplicaciones de aprendizaje automático [Murphy, 2022], identifica los valores de los parámetros que maximizan la probabilidad de los datos empíricos según el modelo asumido [Bishop and Nasrabadi, 2006].

Como se indica en el Apéndice A, Tabla A.1, la *MLE* es una técnica fundamental de inferencia estadística [Casella and Berger, 2024] que abarca diversos enfoques para la estimación de parámetros y la predicción [Bolstad and Curran, 2016]. Por un lado, el enfoque frecuentista asume que los parámetros son constantes desconocidas, siendo la *MLE* un claro ejemplo de este grupo. Por otro lado, el enfoque bayesiano considera que los parámetros son variables aleatorias descritas por distribuciones de probabilidad. La estimación bayesiana combina la información previa (*distribución a priori*) con la evidencia observada (*verosimilitud*) para obtener una *distribución a posteriori* [Gelman et al., 1995]. Ejemplos de técnicas de inferencia bayesiana incluyen la “*Maximum A Posteriori*” (*MAP*) y el “*Markov Chain Monte Carlo*” (*MCMC*), entre otros, como se muestra en el Apéndice A, Tabla A.1. Cabe señalar que, si se asume un prior uniforme, el *MAP* coincide con la *MLE*.

En esta tesis, para determinar los parámetros de las distribuciones teóricas que mejor se ajustan a los datos de distancias recorridas y número de viajes realizados (variables utilizadas para predecir la demanda del servicio), empleamos la biblioteca de Python `SciPy` “*Scientific Python*”, específicamente el submódulo `scipy.stats`, que ofrece un amplio conjunto de herramientas para el análisis estadístico y la manipulación de distribuciones de probabilidad.

Dentro de `scipy.stats` se encuentran diversas distribuciones de probabilidad, tanto continuas como discretas, implementadas como clases que exponen, entre otros, los siguientes métodos:

- `pdf()`: Función de densidad de probabilidad. Se emplea para variables y distribuciones continuas. Mide la densidad de probabilidad en un punto. La integral de la `pdf` sobre un intervalo proporciona la probabilidad de que la variable aleatoria tome un valor dentro de ese intervalo. La `pdf` no se interpreta como una probabilidad directa en un punto específico, de hecho ocurre que $P(X = x) = 0$ (la probabilidad de que la variable tome un valor exacto x es igual a cero porque el “peso probabilístico” se distribuye a lo largo de infinitos puntos).
- `pmf()`: Función de masa de probabilidad. Se emplea para variables y distribuciones discretas. Mide la probabilidad de un valor específico. A diferencia de la `pdf`, que para obtener la probabilidad en un intervalo se integra sobre el mismo, en la `pmf` se suman las probabilidades de los valores discretos que pertenecen al intervalo. También a diferencia de `pdf`, en este caso la probabilidad de que la variable adopte un valor específico k no es cero, sino que está dada por la función de masa de probabilidad (*pmf*) $P(X = k) = p_X(k)$.
- `cdf()`: Se trata de la función de distribución acumulada. Describe la probabilidad acumulada de que la variable sea menor o igual que un valor x que recibe como parámetro. Se emplea indistintamente para funciones continuas como para discretas.

- `fit()`: Se trata de la función de ajuste que permite estimar los parámetros de la distribución (tanto continua como discreta) a partir de datos observados. Cada clase que representa a una distribución diferente, implementa este método el cual inicialmente recibe un `array` de datos. Segudamente aplica un método de optimización, por defecto *MLE*, para encontrar los parámetros que mejor ajusten la distribución teórica a la distribución de los datos aportados. Finalmente el método `fit()` devuelve una lista con los parámetros de la función ajustada. Lógicamente el número de los parámetros dependerá de la función en cuestión.

Para un mayor detalle sobre los parámetros de entrada y los valores de salida de cada una de las funciones mencionadas, se puede consultar la documentación oficial de la biblioteca SciPy, disponible en su [sitio web principal](#) y, de manera más específica, en la sección correspondiente al submódulo `scipy.stats` [SciPy Community, 2024].

En concreto, empleamos la función `fit`, que internamente se basa en el método de Máxima Verosimilitud (MLE). Por consiguiente, se adopta un enfoque frecuentista en el desarrollo de nuestro modelo.

4.2. Modelado de la distancia

Este capítulo detalla el procedimiento llevado a cabo para ajustar los datos empíricos de las distancias recorridas (calculadas conforme a lo expuesto en el capítulo 3.3.3.1 *Método de Cálculo de Distancias en los Viajes de BiciMad*) a la distribución de probabilidad que mejor las representa. Este proceso ha sido documentado en nuestro segundo [Vallez et al., 2024], el cual forma parte integral de esta tesis.

En la literatura se han utilizado diversas distribuciones de probabilidad para modelar las distancias, como se menciona en [Kou and Cai, 2019]. Las distribuciones más utilizadas son la Exponencial, Log-Normal, Rayleigh, Weibull y Gamma.

La distribución exponencial se utiliza para modelar viajes cuando se considera que los eventos de viaje ocurren de forma continua e independiente. Puede interpretarse como el tiempo que le toma a un usuario recorrer una distancia específica bajo la única restricción de que la distancia total (o media) recorrida está fijada Abramovich and Ritov [2022]. La distribución exponencial corresponde a uno de los modelos gravitatorios propuestos para la movilidad interurbana [Bazzani et al., 2010; Liang et al., 2012; Yan et al., 2013]. Por ello, varios estudios la han identificado como la que mejor se ajusta, superando incluso a la de ley de potencias, y ha sido probada tanto en trayectos a pie como en distintos tipos de vehículos. Otros estudios [Rhee et al., 2011; Kölbl and Helbing, 2003] muestran que, tras eliminar los trayectos cortos, la distribución resultante se ajusta mejor a una ley de potencias. En resumen, en la Figura 4.1 se observa que la distribución exponencial logra un ajuste razonable en la cola de la distribución empírica, pero presenta un error considerable en las distancias cortas.

La distribución Log-Normal se emplea para modelar distancias recorridas cuando se considera que estas distancias son producto de varios factores positivos e independientes. Se obtiene mediante una transformación logarítmica de una variable normal. Esta distribución presenta el mismo problema que la anterior. Como se muestra en la Figura 4.1, logra un buen ajuste a la distribución completa, aunque hay desviaciones en los extremos de la distribución empírica. A diferencia de la distribución de ley de potencias, esta ofrece un mejor ajuste cuando se incluyen trayectos cortos en el conjunto de datos, como se ha estudiado en el estudio [Jiang et al., 2009].

La distribución Rayleigh tiene dos variables aleatorias independientes con distribución normal. Esto la hace similar a las distribuciones gaussianas. La distribución Rayleigh es, de hecho, un caso particular de la distribución Weibull cuando el parámetro de forma se establece en 2. Por este motivo, algunos autores prefieren parametrizarla como una Weibull, como se realiza en [Kou and Cai, 2019]. Se observa en la Figura

en la 4.1 que, a pesar de ajustarse mejor al inicio que las anteriores, esta distribución aún presenta un error significativo.

Finalmente, la distribución Gamma se utiliza comúnmente para describir la variación en el tiempo requerido para recorrer una cierta distancia cuando varios factores afectan al trayecto, cambiando las condiciones a lo largo de la ruta.

La Tabla 4.1 presenta un resumen detallado de las distribuciones discutidas anteriormente. En ella se incluye una figura representativa de cada distribución, su función de probabilidad asociada y un breve resumen de sus principales características.

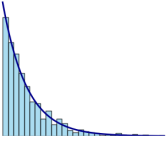
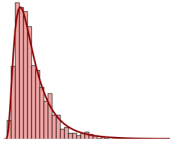
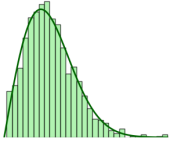
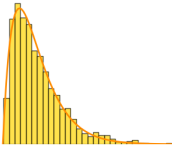
Nombre	Propiedades y Función de Probabilidad
Exponencial 	- Modela viajes independientes con distancia media fija. - Supera a la ley de potencias en movilidad interurbana. - Ajusta bien en la cola, pero falla en distancias cortas. - La función de probabilidad es: $p_d = \lambda e^{-\lambda x}$
Log-Normal 	- Modela distancias con factores positivos e independientes. - Ajusta bien, pero con errores en los extremos. - Mejor que la ley de potencias para trayectos cortos. - La función de probabilidad es: $p_d = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$
Rayleigh/Weibull 	- Modela distancias con dos variables normales independientes. - Es un caso particular de la distribución Weibull con parámetro de forma 2. - Ajusta mejor al inicio, pero aún presenta errores significativos. - La función de probabilidad es: $p_d = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}$
Gamma 	- Modela la variación en el tiempo necesario para recorrer distancias. - Considera múltiples factores que afectan el trayecto. - Captura cambios en las condiciones a lo largo de la ruta. - La función de probabilidad es: $p_d = \frac{\beta^\alpha}{\Gamma(\alpha)} d^{\alpha-1} e^{-\beta d}$

Tabla 4.1: Resumen de las propiedades y de las funciones de probabilidad de las principales distribuciones empleadas en la literatura, para modelar la distancia recorrida en el viaje.

En la Figura 4.1 se muestra el ajuste de los datos empíricos de viajes de BiciMad de 2018 a varias distribu-

ciones utilizadas en la literatura para modelar la distancia recorrida. Se presenta tanto en escala logarítmica como escala normal. En el gráfico se destacan tres secciones diferentes:

- Sección 1: Representa los viajes de corta distancia. En este grupo de trayectos, todas las distribuciones fallan al intentar ajustarse a los datos empíricos, aunque la distribución Gamma es la que más se aproxima.
- Sección 2: Representa los trayectos más comunes. En este caso, todas las distribuciones parecen ajustarse bien, siendo la Normal la que peor se ajusta desde un punto de vista visual.
- Sección 3: Representa los trayectos largos. Aquí, nuevamente, todas las distribuciones tienen problemas para ajustarse adecuadamente. Sin embargo, visualmente, la distribución Gamma parece ser la que mejor aproxima los valores empíricos.

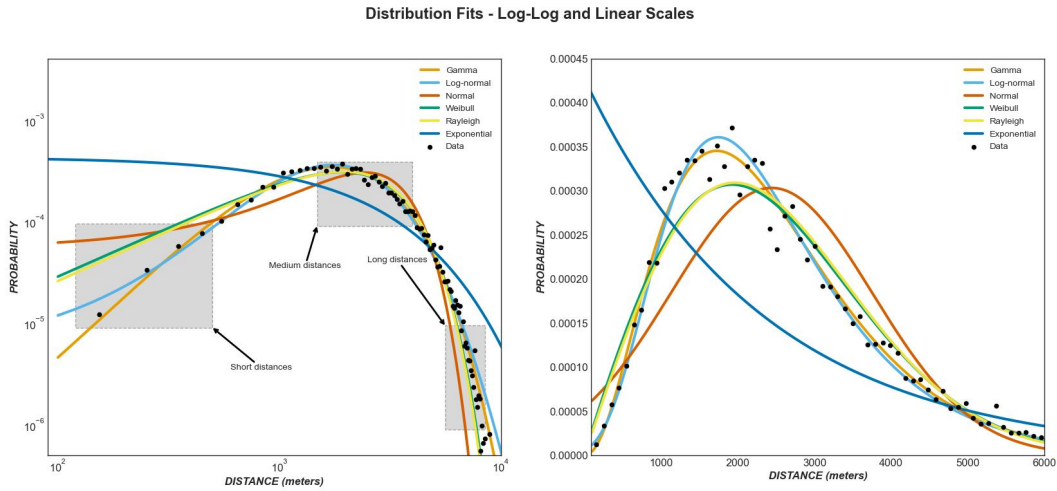


Figura 4.1: Distribución empírica de viajes entre estaciones a una distancia d para el conjunto de datos de BiciMad de 2018. A la izquierda: ajuste a las distribuciones de probabilidad más comunes utilizadas para modelar distancias en escala doblemente logarítmica; a la derecha: el mismo gráfico en escala lineal. Obsérvese cómo esta representación (típica) puede ser engañosa, ya que no proporciona suficiente resolución en las colas. En general, la distribución Gamma es la que mejor captura todos los puntos de datos del conjunto, seguida de cerca por la distribución log-normal. En este capítulo se realiza una comparación más cuantitativa. Fuente: Datos de viajes de Bicimad 2018 y gráfico de elaboración propia.

Se concluye que, desde un punto de vista visual, la distribución Gamma parece ofrecer el mejor ajuste, aunque será validado utilizando el método de Kullback-Leibler en el siguiente subcapítulo.

4.2.1. Validación del ajuste de la distribución mediante test de Kullback-Leibler

Este capítulo está dedicado al uso del test de Kullback-Leibler para evaluar cuál de las distribuciones estadísticas propuestas se ajusta con mayor precisión a la distribución empírica de las distancias de los viajes realizados en 2018 en el sistema BiciMad.

Desde una perspectiva estrictamente física, las posibles distancias deberían corresponder únicamente a los valores finitos del conjunto de combinaciones origen-destino entre estaciones, considerándose como

datos discretos. Sin embargo, teniendo en cuenta los contextos presentados en la Tabla 4.2, se puede tratar a los datos de distancias como continuos.

Contexto	Explicación
Variaciones en las trayectorias	Para modelar el origen-destino de un viaje se ha decidido asignar una distancia concreta al viaje, según se ha explicado en capítulo X. Sin embargo las trayectorias prácticas de los usuarios varían, ya que deciden tomar rutas alternativas (debido al tráfico o a obras por ejemplo). También se constata que a veces las trayectorias seguidas por los usuarios son contrarias a las normas de circulación, bajando de la bicicleta por ejemplo y haciendo los últimos metros hasta la estación por la acera y en sentido contrario al de la circulación de la calle.
Diferentes distancias según el sentido	Si el viaje va de la estación A a la B, tiene una distancia diferente a si va de la estación B a la A. Aunque la diferencia es mínima. En el capítulo X se explicó el sentido de esta decisión.
Conjunto variable de estaciones	Otra razón para considerar las distancias como datos continuos es que, aunque en un momento dado solo exista un conjunto limitado de distancias predefinidas entre estaciones, la realidad es que este conjunto puede variar con el tiempo. Año tras año, la incorporación de nuevas estaciones genera distancias que anteriormente no eran posibles, mientras que eventos externos, como el cierre temporal o definitivo de estaciones, o la imposibilidad de tomar o dejar bicicletas en puntos específicos, pueden eliminar distancias previamente existentes. Esto refleja una dinámica que, si bien es discreta en un instante dado, a su vez es potencialmente continua en el largo plazo debido a la evolución del sistema y sus configuraciones.
Muestras de tamaños grande	Cuando los tamaños de muestra son suficientemente grandes, muchas distribuciones discretas muestran un comportamiento cercano al de una distribución continua. Esta idea ha sido enunciada de forma particular en el Teorema Central del Límite (CLT), el cual establece que la suma de variables aleatorias independientes e idénticamente distribuidas, con media y varianza finitas, converge a una distribución normal a medida que el tamaño de la muestra aumenta. Por lo tanto, aun teniendo datos empíricos que provienen de una distribución discreta, el comportamiento de la media muestral puede aproximarse por una distribución continua normal para tamaños de muestra grandes.
Aproximaciones en modelado	En estadística, es común modelar datos discretos utilizando distribuciones continuas para simplificar el análisis y aprovechar las propiedades matemáticas de estas distribuciones. Por ejemplo, en el libro “ <i>Pattern Recognition and Machine Learning</i> ” [Bishop and Nasrabadi, 2006], concretamente en los capítulos sobre modelos híbridos y métodos de variables latentes ¹ , se plantea el uso de espacios latentes ² continuos para modelar datos discretos. Otro ejemplo se encuentra en [Hastie, 2009] donde se explora enfoques de modelado paramétrico y no paramétrico, incluyendo casos en los que se utilizan funciones continuas, para aproximar datos categóricos (discretos) en el modelado predictivo.

Tabla 4.2: Contextos que justifican el tratamiento de las distancias como datos continuos en lugar de discretos.

Por lo tanto, se puede concluir que, en el caso de BiciMad, el comportamiento agregado de las distancias recorridas en múltiples viajes puede aproximarse a una distribución continua debido al tamaño de la muestra,

¹Una variable latente es una variable no observable directamente que explica las relaciones entre las variables observadas en los datos.

²Un espacio latente es una representación matemática que captura patrones ocultos en los datos de forma simplificada y organizada en una dimensión más reducida que los datos originales.

lo cual es consistente con el *CLT*. Adicionalmente hay que señalar que los datos de distancias de BiciMad recogen viajes largos poco frecuentes, lo cual se traduce en colas en la distribución.

La tabla 4.3 muestra diversas alternativas al test de *Kullback-Leibler*, destacando las limitaciones asociadas a cada una de ellas, lo que justificó nuestra decisión de seleccionar dicha métrica. En esta tabla se han excluido métricas que aparecen en la literatura y que tienen poca interpretación estadística o no directa.

Se concluye por tanto que al tratar las distancias de los viajes en BiciMad como datos continuos y tener colas significativas (viajes largos poco frecuentes), *KL* es la métrica adecuada en este caso para evaluar la diferencia entre la distribución empírica y la probabilística teórica ajustada. Esta métrica detecta diferencias no solo en las regiones en torno a la moda, sino también en las colas, lo que es crucial para una representación estadística precisa.

Métrica	Descripción / Limitaciones
Kolmogorov-Smirnov (<i>KS</i>)	Mide la diferencia máxima entre las funciones de distribución acumulada (<i>CDF</i>) de dos distribuciones. Por tanto no considera la diferencia en toda la distribución, siendo menos sensible en zonas concretas donde no se da la diferencia máxima como es el caso de las colas.
Media Absoluta del Error (<i>MAE</i>)	Se trata de una métrica sencilla que mide las diferencias punto a punto, al contrario de <i>KL</i> que lo hace para la distribución completa. Como desventaja hay que destacar que penaliza todos los errores por igual, no teniendo en cuenta que por ejemplo en nuestro caso hay viajes más frecuentes que otros.
Raíz del Error Cuadrático Medio (<i>RMSE</i>)	Se trata de otra métrica sencilla punto a punto, la cual penaliza más fuerte los errores más grande siendo por tanto sensible a los outliers.
Divergencia de Chi Cuadrado χ^2	Se utiliza principalmente para comparar frecuencias observadas con frecuencias esperadas en datos categóricos. Aunque puede adaptarse para datos continuos mediante la agrupación en categorías (<i>binned chi-square test</i>)
Divergencia de Jensen-Shannon (<i>JSD</i>)	Es una variante simétrica ³ de <i>KL</i> . Esto significa que no depende del orden en el que se comparan las distribuciones, porque toma un promedio ponderado de las divergencias <i>KL</i> . Adicionalmente, hay que resaltar que es más costosa de calcular en términos de computación

Tabla 4.3: Descripción y limitaciones de métricas alternativas al test de *Kullback-Leibler*

En teoría de la información, la divergencia de *Kullback-Leibler* (*KL*), también conocida como entropía relativa, mide la diferencia entre dos distribuciones de probabilidad. Sirve para calcular la “distancia” entre estas distribuciones, según lo señalado en [Pérez-Cruz, 2008], indicando la cantidad de información que se pierde al aproximar una distribución con otra.

Para distribuciones de probabilidad discretas, la fórmula de la divergencia *KL* es:

$$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right),$$

³Matemáticamente, la simetría de *JSD* se expresa como: $JSD(P \parallel Q) = JSD(Q \parallel P)$ mientras que la asimetría de *KL* se indica como: $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$.

donde P y Q son las distribuciones de probabilidad, x son los diferentes puntos que se evalúan en dichas distribuciones, y X es el espacio en el cual x está definido. En el caso continuo, se reemplaza el sumatorio por una integral, pero numéricamente tratamos las distribuciones como discretas.

En la implementación de la fórmula en el código, se utiliza un pequeño parámetro ϵ para evitar código condicional que verifique que ni P ni Q sean iguales a 0 (básicamente para evitar divisiones por cero e indeterminaciones como $\log(0)$). Este parámetro ϵ es lo suficientemente pequeño como para no afectar el resto de los valores y simplemente se establece por conveniencia.

$$D_{KL}(P \parallel Q) = \sum_{x \in X} (P(x) + \epsilon) \log \left(\frac{(P(x) + \epsilon)}{(Q(x) + \epsilon)} \right) \quad \text{with } \epsilon = 1e-05$$

Una vez establecida esta metodología para calcular D_{KL} , los resultados obtenidos permiten comparar cómo diferentes distribuciones teóricas se ajustan a los datos empíricos. En este contexto, el uso de herramientas como la biblioteca `scipy.stats` de Python facilita la definición de distribuciones continuas. En dicha biblioteca, las distribuciones continuas generalmente se definen con dos parámetros: `shape` y `scale`. El parámetro `shape` controla la asimetría de la distribución, mientras que `scale` estira o comprime la distribución. Sin embargo, la interpretación puede tener un significado especial en algunos casos. La Tabla 4.4 presenta un análisis comparativo de los parámetros y la divergencia de *Kullback-Leibler* (KL) entre los datos empíricos de viajes de BiciMad de 2018 y las distribuciones teóricas consideradas en el proceso de ajuste. La divergencia KL se aplicó a la distribución de todos los viajes sin realizar subagrupaciones adicionales (mes, semana, día, hora, etc.).

Distribución	Parámetros obtenidos en el ajuste	Valor D_{KL}
Gamma	Shape: 3.44	4.5e-05
	Scale: 714.58	
Lognormal	Shape: 0.40	7.9e-05
	Scale: 2975.81	
Weibull	Shape: 1.97	1.5e-04
	Scale: 2761.65	
Rayleigh	Shape: valor fijo de 2 porque es un caso particular de la distribución Weibull.	1.5e-04
	Scale: 1960.28	
Normal	Media: 2440.80	8.6e-04
	Desviación estándar (Scale): 1314.51	
Exponential	Scale: 2336.19.	2.4e-03
	Siendo $\lambda = 1/\text{Scale}$ aprox 4.3e-04	

Tabla 4.4: Parámetros y valores de la Divergencia KL para las distribuciones consideradas, y ordenadas de mejor (menor KL) a peor ajuste (mayor KL).

Como se muestra en la Tabla 4.4, la distribución Gamma tiene el valor más bajo de divergencia (KL) con respecto a los datos. En segundo lugar, la distribución Lognormal también muestra un ajuste muy bueno y podría ser una alternativa relativamente aceptable a la Gamma. Como se observa en la Figura 4.1, las diferencias son leves y solo afectan a unos pocos puntos de datos en los extremos. En contraste, las distribuciones Weibull y Rayleigh capturan algunos aspectos de los datos empíricos, pero con menos precisión que las anteriores. Hay que recordar que la distribución Rayleigh es, de hecho, un caso particular de la distribución

Weibull. Finalmente, la distribución Normal no puede predecir la asimetría (y la positividad de los datos), mientras que la distribución Exponencial tiene su moda en la distancia cero, por lo que es claramente incapaz de capturar el pico de la distribución alrededor de los 2000 m. De ahí que estas otras distribuciones tengan un valor de KL superior al de la Gamma.

A falta de una teoría de primeros principios más satisfactoria (más allá de los argumentos basados en el principio de máxima entropía que discutimos más abajo en este capítulo), se concluye que la distribución Gamma es la que mejor se ajusta para modelar la distancia de los viajes de BiciMad, tanto de forma visual, como se observa en la Figura 4.1, como de manera cuantitativa, según los resultados del test de KL .

4.2.2. Ajuste de los datos empíricos a la distribución Gamma

Después de haber seleccionado la distribución Gamma como la que mejor captura todas las distancias en los datos, se ajustan los datos empíricos a dicha distribución pero con granularidad horaria. Esto tiene el beneficio de que nuestro simulador (ver capítulo 4.4) pueda capturar las diferencias horarias en la demanda. Además, en el capítulo 4.3 se aprecia que existen grandes diferencias en el número de viajes por hora en distintos momentos del día.

Para la parametrización de la distribución Gamma se emplean los valores α y $scale = 1/\beta$. La Figura 4.2 (parte inferior) muestra los valores de los parámetros de forma y escala para las distribuciones Gamma ajustadas a los viajes empíricos. En el parte superior de dicha figura, se observa que, aunque los parámetros presentan cierta variación, la distribución general es bastante similar para todas las horas. Adicionalmente en la Tabla 4.5 se reflejan los valores obtenidos por hora para los parámetros α y $scale$ tras el ajuste de los datos empíricos a la distribución Gamma.

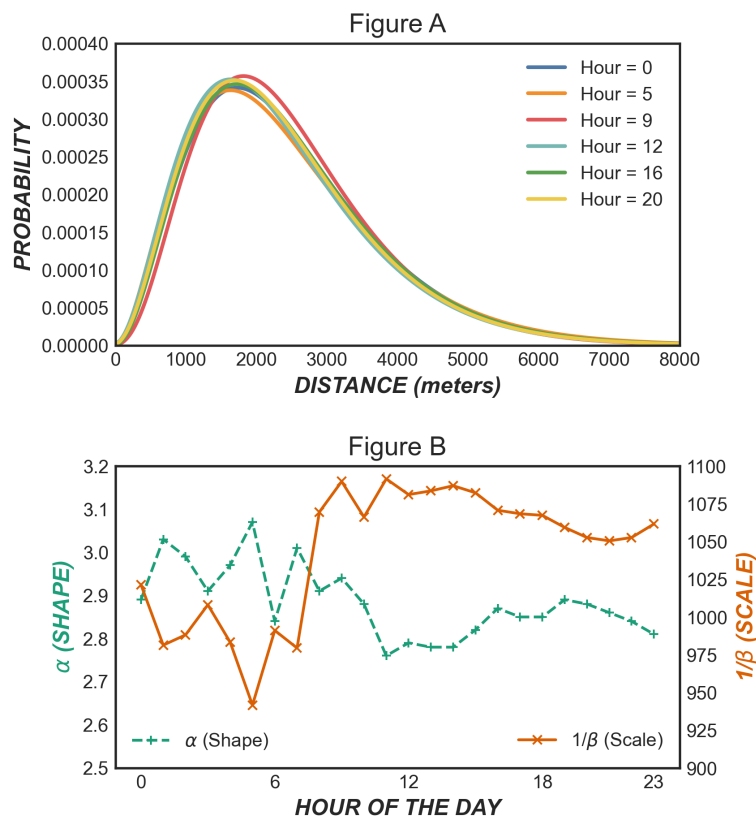


Figura 4.2: Parte superior: Comparación de las distribuciones Gamma de mejor ajuste para una muestra de horas representativas. Aunque los parámetros en el panel inferior muestran cierta variación a lo largo del día, la forma de la distribución apenas cambia. Parte inferior: Valores de los parámetros *shape* y *scale* de la distribución Gamma ajustada para diferentes horas del día. El parámetro *scale* está relacionado con una distancia media típica de viaje, por lo que el panel inferior sugiere que dicha distancia es mayor durante las horas punta. Fuente: Elaboración propia.

Hora	0	1	2	3	4	5	6	7	8	9	10	11
Shape: α	2.89	3.03	2.99	2.91	2.97	3.07	2.84	3.01	2.91	2.94	2.88	2.76
Scale: θ (metros)	1021.45	981.35	988.14	1008.06	983.28	941.62	991.08	979.43	1069.52	1089.32	1066.10	1091.70
Hora	12	13	14	15	16	17	18	19	20	21	22	23
Shape: α	2.80	2.78	2.78	2.82	2.87	2.85	2.85	2.89	2.88	2.86	2.84	2.81
Scale: θ (metros)	1081.08	1083.42	1086.96	1082.25	1070.76	1068.38	1067.24	1059.32	1052.63	1050.42	1052.63	1061.57

Tabla 4.5: Listado de parámetros *Shape* (α , adimensional) y *Scale* (θ , medido en metros) de las funciones de distribución Gamma obtenidas para cada hora del servicio.

4.2.3. Justificación de la idoneidad de la distribución Gamma

Para representar las distancias de los viajes, se probaron varias distribuciones de probabilidad comúnmente utilizadas en la literatura como son la Exponencial, Lognormal, Weibull/Rayleigh y la propia Gamma. Se realizó una validación visual comparando gráficamente cuál se ajustaba mejor a los datos empíricos y se midió “cuánto se pierde” o lo que es lo mismo “la similitud” entre los datos reales y las distribuciones teóricas. El resultado fue que la distribución Gamma presentó el mejor ajuste.

En este punto surge la pregunta ¿Por qué la distribución Gamma presenta un ajuste bastante preciso a los datos?

Primeramente hay que señalar que la distribución Gamma se utiliza frecuentemente en situaciones donde existen múltiples factores que afectan el resultado de un evento, las cuales generan variaciones del mismo. En el caso de BiciMad la distancia de los viajes se encuentra condicionada por varios factores como son el tráfico, la disponibilidad de bicicletas en estación origen y de anclajes en estación destino, la hora, el clima y factores demográficos, económicos y sociológicos entre otros. Con el parámetro de forma α que controla la asimetría y el parámetro $\Theta = Scale$ que gestiona la dispersión de los datos, la distribución Gamma presenta su flexibilidad característica para adaptarse a as diversas situaciones.

Sin embargo, no está del todo claro por qué la Gamma ajusta tan bien los datos a lo largo de todo el rango de distancias. Se considera que ello puede deber a dos motivos:

- Principio de Máxima Entropía: Una posible explicación es que la agregación de datos incompletos sobre usuarios individuales (sus motivaciones, frecuencia de uso, restricciones económicas, etc.) disipa cualquier información específica, lo que sugiere que una caracterización basada en el principio de máxima entropía puede capturar los datos [Jaynes, 2003]. Este principio subyace en el modelo gravitacional (Anderson, 2011), ampliamente utilizado en el contexto del modelado de matrices de origen-destino (O-D). Según esta suposición, se necesita maximizar la entropía de Shannon⁴, considerando las restricciones relevantes al problema (verificando si la distribución resultante describe con precisión los datos). Aquí, se aplica el principio de máxima entropía considerando dos restricciones: la primera está vinculada a la distancia promedio $\bar{d}$ y la segunda a su logaritmo. Esto se basa en la suposición de que la percepción humana sobre cantidades, como el dinero, el tiempo, o las distancias, entre otras, suele describirse mejor en esa escala. Esta idea de la percepción logarítmica se conoce como la ley de Weber-Fechner de percepción logarítmica [Dehaene, 2003]. Por lo tanto, la segunda restricción es: $\bar{L} \equiv \langle \log d \rangle$. De este modo, al aplicar la maximización utilizando multiplicadores de Lagrange β y $\alpha - 1$ (por conveniencia), obtenemos:

$$S = - \sum p_d \log p_d - \lambda \left(\sum p_d - 1 \right) - \beta \left(\sum p_d d - \bar{d} \right) - (\alpha - 1) \left(\sum p_d \log d - \bar{L} \right),$$

Maximizarla significa encontrar la distribución más imparcial, o que es lo mismo menos sesgada, que cumpla con las restricciones que son:

- Restricción de normalización: La suma de todas las probabilidades debe ser 1.

$$\sum p_d = 1$$

Esto asegura que p_d sea una distribución de probabilidad válida. Su multiplicador asociado es λ

- Restricción de la distancia promedio: La distancia promedio debe ser igual a $\bar{d}$.

$$\sum p_d d = \bar{d}$$

Esto asegura que la distribución cumpla con el promedio observado. Su multiplicador asociado es β

- Restricción del logaritmo de la distancia promedio: El logaritmo de la distancia promedio debe ser igual a $\bar{L}$.

$$\sum p_d \log d = \bar{L}$$

Esto se vincula con la percepción logarítmica (Ley de Weber-Fechner). Su multiplicador asociado es $\alpha - 1$. Se emplea esta notación porque facilita la escritura de las ecuaciones y su posterior resolución.

El primer término de la ecuación corresponde a la entropía de Shanon que mide la incertidumbre asociada a una distribución p_d . Matemáticamente⁵ se expresa como:

$$H(p_d) = - \sum p_d \log p_d$$

Siguiendo con el cálculo, se deriva S respecto a p_d y luego igualar a 0

$$\frac{\partial S}{\partial p_d} = 0$$

y se obtiene que la forma de p_d debe ser:

$$p_d \propto d^{\alpha-1} e^{-\beta d}$$

Esto significa que p_d es proporcional $d^{\alpha-1} e^{-\beta d}$.

Para que sea una distribución de probabilidad válida y por tanto la suma total de las probabilidades sea 1, es necesario normalizar usando en este caso la constante de normalización: $\frac{\beta^\alpha}{\Gamma(\alpha)}$.

La solución es la fórmula de la Distribución Gamma:

$$p_d = \frac{\beta^\alpha}{\Gamma(\alpha)} d^{\alpha-1} e^{-\beta d}$$

El significado práctico de esta fórmula es que para valores pequeños de d se da un crecimiento inicial controlado por el parámetro α mientras que para los valores grandes de d se produce una caída exponencial controlada por el parámetro $\beta = 1/\theta$ que es el inverso del parámetro *Scale* que se emplea en esta tesis para parametrizar la Gamma.

- Interpretación de la Distribución Gamma como suma de variables exponenciales idénticas: Otra forma en la que podría surgir la distribución Gamma es cuando el parámetro de forma α es un número entero, caso en el que la distribución resulta ser la suma de un número entero de distribuciones exponenciales idénticas. Esto correspondería a la distribución de Boltzmann bajo una distancia total constante. Sin embargo, como el parámetro α no es entero y en todos los casos es menor que 3, es decir, menos de tres pasos idénticos, esta explicación no resulta satisfactoria.

⁴La entropía mide el “nivel de incertidumbre” o la “cantidad de información” que contiene una distribución de probabilidades. Así por ejemplo una moneda no trucada $p(\text{cara}) = 0,5$ y $p(\text{cruz}) = 0,5$ tiene máxima entropía porque no se puede predecir qué saldrá. En cambio una moneda trucada que siempre sale cara $p(\text{cara}) = 1$ y $p(\text{cruz}) = 0$ tiene entropía = 0 porque no hay incertidumbre.

⁵La fórmula incluye un signo negativo porque, aunque la entropía de Shannon, es positiva o igual a cero por definición, se busca convertir un problema de minimización en uno de maximización.

De manera resumida, se concluye que la distribución Gamma es adecuada para modelar las distancias de los viajes del servicio de BiciMad, debido a su capacidad para capturar la variabilidad asociada a diversos factores. Su ajuste se explica principalmente por el Principio de Máxima Entropía, derivado de restricciones como la distancia promedio y la percepción logarítmica. Otras posibles explicaciones, como la suma de distribuciones exponenciales idénticas, no es aplicable en este caso debido a que los parámetros α de las distintas distribuciones Gamma a nivel de hora no son números enteros.

4.3. Modelado del número de viajes

Siguiendo un procedimiento análogo al de la sección anterior, en esta sección se categoriza el número de viajes por hora, dado que el servicio se utiliza tanto en horas punta como en horas valle. Adicionalmente, la información disponible tiene una granularidad temporal mínima de una hora, sin datos disponibles a nivel de minutos. Por lo tanto, en el modelo, un viaje que comience a las 9:15 o a las 9:35 se asignará al grupo de las 9:00. El objetivo es obtener una distribución teórica que prediga el número de viajes que ocurren en un marco temporal de una hora. Posteriormente, esto permitirá simular el origen y el destino de cada uno de estos viajes a nivel probabilístico utilizando la distribución Gamma descrita anteriormente.

En una primera aproximación a este ajuste, sólo se emplearon las horas en esta categorización. Como resultado, en algunos casos fue posible ajustar una distribución binomial negativa (*NegBin*), mientras que en otros casos surgió ruido en el ajuste.

La distribución binomial negativa representa el número de intentos independientes e idénticos necesarios para alcanzar un número determinado de éxitos. Su función de masa de probabilidad es:

$$f(k) = \binom{k+n-1}{n-1} \cdot p^n \cdot (1-p)^k$$

Donde n representa el número de éxitos y p es la probabilidad de un único éxito.

Otra forma de parametrizar la distribución binomial negativa, que es la empleada en esta tesis, se expresa en términos de la media μ y la varianza σ^2 :

$$p = \frac{\mu}{\sigma^2} \quad n = \frac{\mu^2}{\sigma^2 - \mu}$$

Tras un examen detallado de los días que incluían ruido en la distribución, se descubrió la existencia de dos subpatrones distintos dentro de las horas, dependiendo de si correspondían a días laborales o fines de semana. También se consiguió determinar que los días festivos, a efectos del modelo, se comportaban como fines de semana.

A pesar de esta división dentro de las horas según el tipo de día, el ajuste a la distribución binomial negativa seguía presentando algunas anomalías. Se decidió explorar el impacto de los factores meteorológicos. Las condiciones climáticas, particularmente la precipitación, influyen significativamente en la decisión de realizar un viaje en un momento específico. En cuanto a la temperatura, se acordó no incluirla en el modelo por dos razones: en primer lugar, como se menciona en la Ref. [del Río et al., 2011], las regiones occidentales y centrales de la Península Ibérica muestran un patrón claro, donde los inviernos más fríos son los más lluviosos y los veranos más cálidos, lo que sugiere una correlación entre la temperatura y la precipitación. En segundo lugar, esta decisión también se respaldó en el análisis realizado sobre el número de viajes diarios frente a las temperaturas máxima, mínima y promedio del día, indicando adicionalmente si el día era festivo/fin de semana o laboral, así como si hubo o no precipitación.

En el Apéndice F la Figura F.1 presenta de manera visual el análisis mencionado anteriormente. Adicionalmente, La Figura F.2 muestra los resultados obtenidos del Análisis de Regresión Lineal de Viajes Diarios

en función de las Variables Meteorológicas y del Tipo de Día. Basándonos en el Análisis de Regresión, se han observado las siguientes conclusiones:

- Durante los fines de semana o días festivos, hay, en promedio, 3064 viajes menos en comparación con los días laborales.
- Por cada grado adicional en la temperatura promedio, el número de viajes aumenta aproximadamente en 174, en promedio. Considerando un promedio de 9128 viajes diarios con una desviación estándar de 3184 viajes, un incremento o decremento de 174 viajes no resulta significativo.
- En los días con precipitación, hay, en promedio, 1158 viajes menos en comparación con los días sin precipitación.

Como resultado, se definen dos nuevas variables para cada viaje:

- `weekend_holiday_flag`: distingue los viajes realizados durante los fines de semana o días festivos (valor=1) de aquellos realizados en días laborales (valor=0). Para obtener esta variable, se realiza una unión con el conjunto de datos de festivos por día.
- `rain_flag`: indica si la precipitación es igual o mayor a 1 mm (valor=1; en caso contrario, valor=0). Se eligió el valor de 1 mm porque representa 1 litro de lluvia en 1 metro cuadrado y se encuentra en el centro de la banda de intensidad de lluvia "débil".

A continuación, se crea una familia de distribuciones binomiales negativas (*NegBin*) para cada hora y combinación de escenarios (lluvioso o no lluvioso, día laboral o fin de semana-festivo) utilizando la función `fits()` del paquete de Python *scipy.stats* [SciPy Community, 2024]. La Figura 4.3 muestra cómo los valores medios y las desviaciones estándar del número de viajes siguen patrones similares en diferentes escenarios (tipo de día y nivel de precipitación), aunque en una escala diferente.

Las distribuciones *NegBin* generadas para las diferentes horas en los escenarios de días laborales sin lluvia se presentan en la Figura 4.4. Nótese cómo el ajuste sigue con gran precisión las distribuciones empíricas subyacentes. Una posible razón detrás de esta alta precisión radica en que la distribución *NegBin* puede interpretarse como una mezcla (hipotéticamente relacionada con preferencias individuales) de distribuciones Poisson, siendo estas últimas la distribución de máxima entropía para conteos. La Tabla 4.6 presenta un resumen que permite al lector hacerse una idea de la estructura y los valores generales de los parámetros p y n de la distribución *NegBin*, mientras que el detalle completo puede consultarse en el apéndice G.

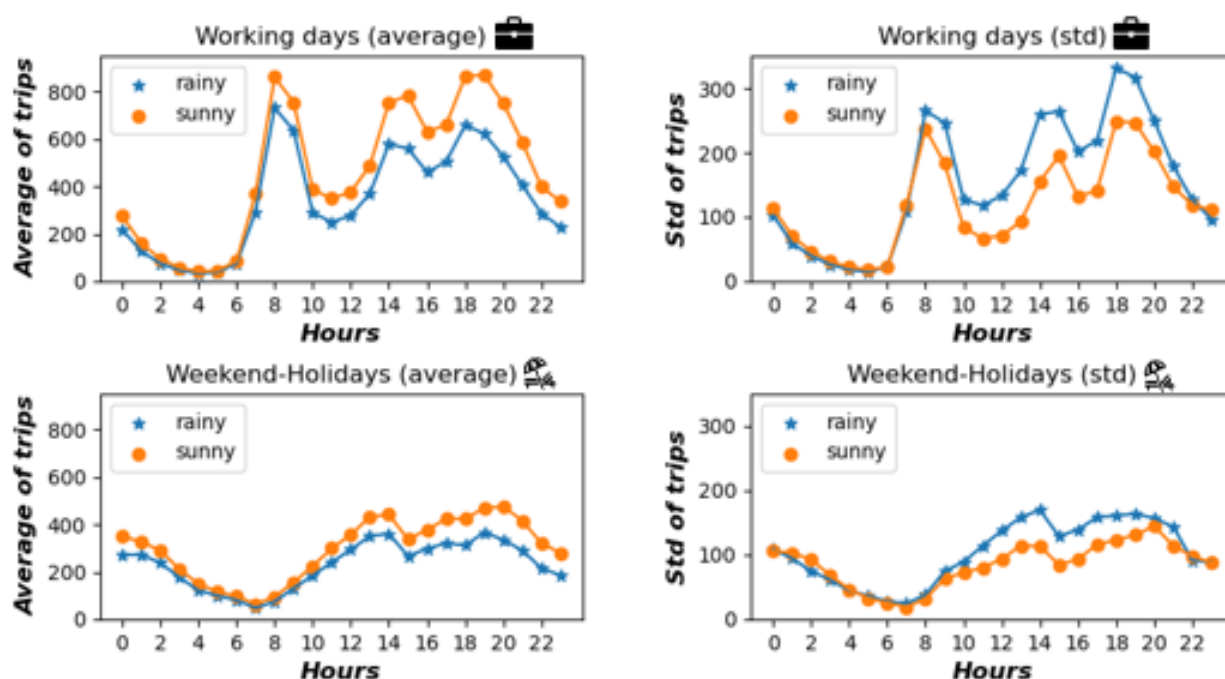


Figura 4.3: Promedio (columna izquierda) y desviación estándar (columna derecha) del número de viajes por hora bajo diferentes condiciones. En general, el promedio es menor en los días lluviosos y su variabilidad es mayor. El efecto de los fines de semana o días festivos escala todas las curvas y modifica el patrón de "horas punta" entre las 8-19 y 14-18, típico de los días laborables. Fuente: Datos extraídos de los viajes de 2018 de BiciMad y gráfico de elaboración propia

weekend_holiday_flag	rain_flag	hour	p	n	average	var	std
0	0	0	0.02	6.05	275.95	12864.55	113.42
0	0	1	0.03	5.33	161.05	5025.52	70.89
0	0	2	0.04	4.46	95.14	2126.61	46.12
...	...	...	...	...	...	...	...
0	1	6	0.14	12.23	74.00	521.63	22.84
0	1	7	0.02	7.41	291.02	11725.02	108.28
...	...	...	...	...	...	...	...
1	0	15	0.05	16.70	336.98	7137.17	84.48
1	0	16	0.04	17.09	377.62	8719.89	93.38
...	...	...	...	...	...	...	...
1	1	21	0.01	4.14	290.46	20668.82	143.77
1	1	22	0.03	5.70	215.43	8358.46	91.42
1	1	23	0.02	4.44	185.39	7932.24	89.06

Tabla 4.6: Subconjunto ejemplo de los parámetros (p y n) de la distribución NegBin para diferentes horas y clasificados por tipo de día y condiciones de precipitación.

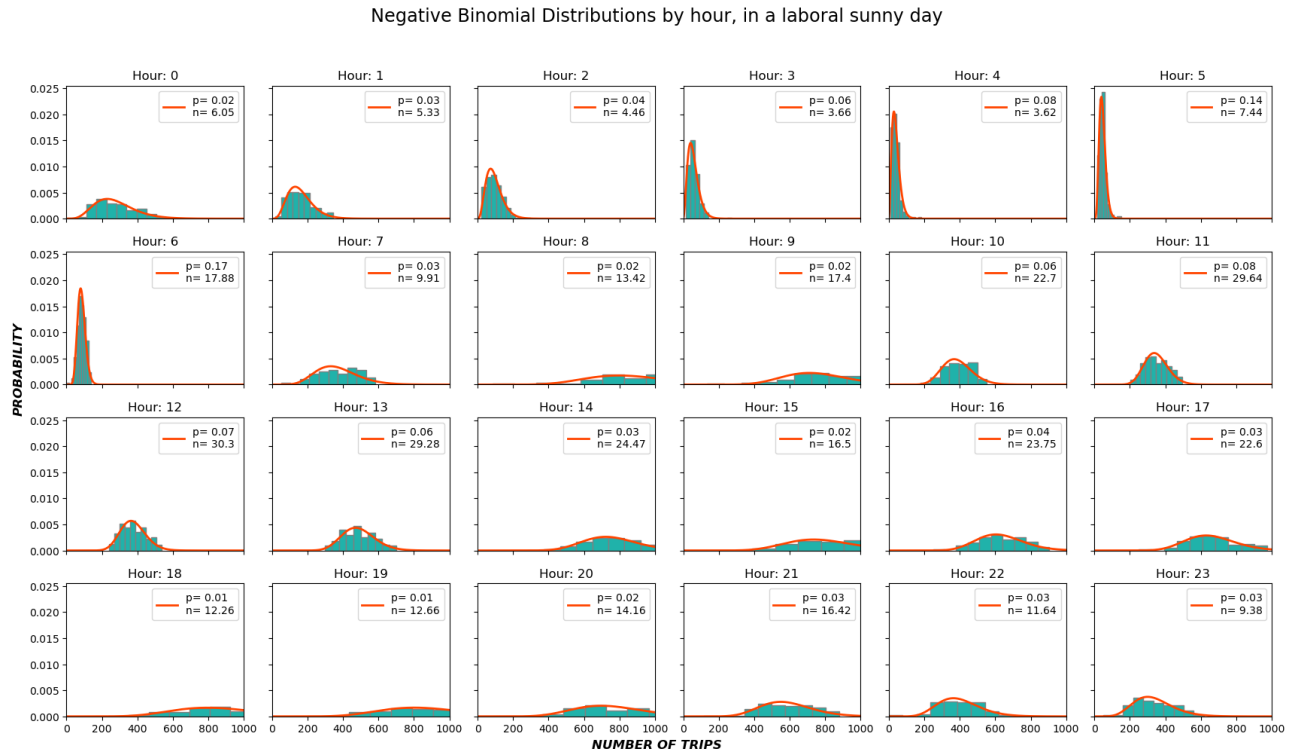


Figura 4.4: Parámetros y gráficos de la distribución NegBin para días soleados y laborables por hora. Nótese la precisión con la que la distribución se ajusta a los histogramas empíricos subyacentes.

4.3.1. Generación de datos sintéticos y validación interna

En los capítulos anteriores se han estimado los parámetros para las distribuciones Gamma y NegBin, que explican las distancias y los números de viajes del servicio BiciMad. Utilizando estos parámetros, se pueden generar datos sintéticos muestreando a partir de dichas distribuciones. Esto es similar a lo que, en la terminología de la inferencia Bayesiana, se conoce como Posterior Predictive Checks. Esto permite generar réplicas de datos sintéticos aplicando el modelo ajustado del año 2018 para ambas distribuciones y compararlas con los datos reales observados en 2019. Antes de llevar a cabo la validación externa con los datos de 2019 en el Capítulo 5.2, este capítulo se centra en realizar una validación interna del número de viajes generados a partir de las distribuciones binomiales obtenidas.

Para realizar la validación interna y utilizando los mismos datos de 2018 empleados para ajustar las distribuciones, se construye un conjunto de datos a nivel horario sujeto a diferentes condiciones (fines de semana/días festivos y si llueve o no) utilizando distribuciones de probabilidad binomial negativa para calcular intervalos de confianza y estadísticas relevantes. Usando las funciones de Python para la distribución binomial negativa (`nbinom.ppf`), calculamos los límites inferior y superior del intervalo de confianza al 95 % (percentiles del 2.5 % y 97.5 %). El número de valores empíricos del conjunto de datos de 2018 por debajo del límite inferior o por encima del límite superior se cuenta para cada combinación de condiciones. Además, se calcula el porcentaje de valores que caen dentro del intervalo de confianza como:

Para realizar la validación interna, utilizando los mismos datos de 2018 empleados para ajustar las distribuciones, se construye un conjunto de datos a nivel horario sujeto a diferentes condiciones (fines de semana/días festivos y si llueve o no). Se emplean distribuciones de probabilidad binomial negativa para calcular intervalos de confianza y estadísticas relevantes. Usando las funciones de Python para la distribución bino-

mial negativa (*nbinom.ppf*), se determinan los límites inferior y superior del intervalo de confianza al 95 % (percentiles del 2.5 % y 97.5 %). El número de valores empíricos del conjunto de datos de 2018 por debajo del límite inferior o por encima del límite superior se cuenta para cada combinación de condiciones. Además, se calcula el porcentaje de valores que se encuentran dentro del intervalo de confianza como:

$$\text{percentage} = \frac{\text{total} - (\text{numhi} + \text{numlow})}{\text{total}} \times 100$$

Con la información obtenida, que se puede consultar en el Apéndice H, se generó la Figura 4.5, la cual muestra los intervalos del número de viajes en función de la hora del día para cada uno de los cuatro escenarios (días laborables o festivos frente a condiciones de lluvia o sin lluvia).

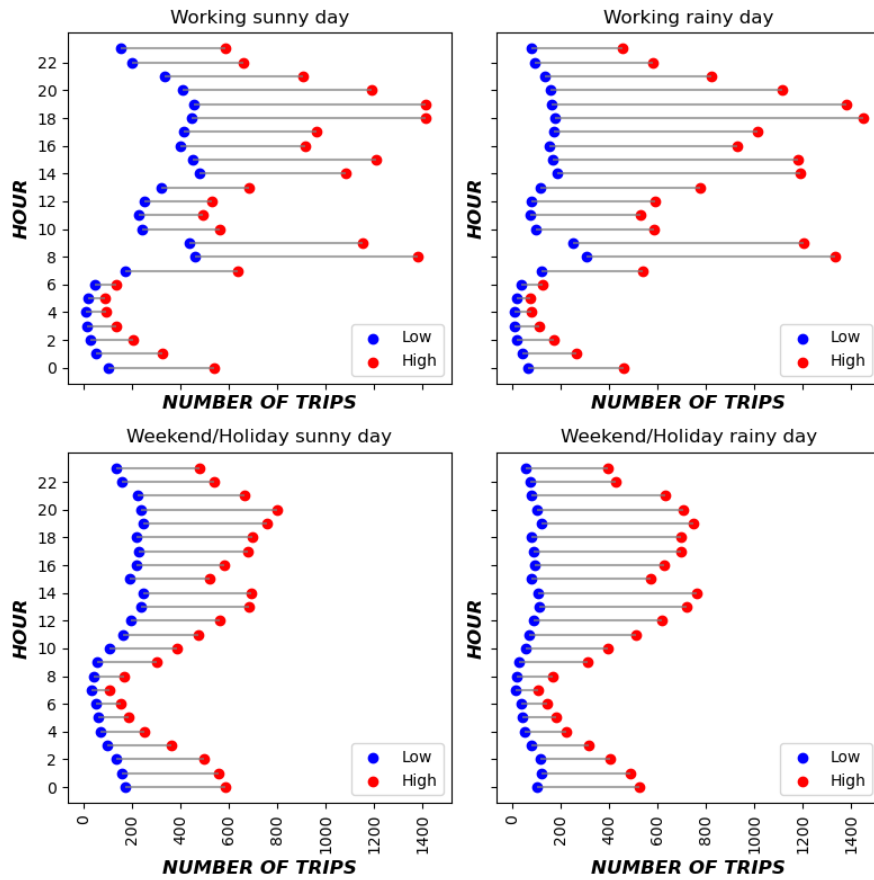


Figura 4.5: Intervalo (percentiles 2,5 % - 97,5 %) para el número de viajes por hora y tipo de día, obtenido a partir de distribuciones Negbin. Fuente: Elaboración propia

Finalmente, se resumió la información de los intervalos, obteniendo los conteos totales y los porcentajes calculados para cada hora. Se pudo verificar que los valores están alrededor del 95

4.4. Creación de un simulador probabilístico

Con las distribuciones previamente establecidas, se construyó un simulador probabilístico integrando dichas distribuciones con información adicional sobre la popularidad de cada estación. Este simulador per-

mite responder preguntas causales del tipo “¿qué pasaría si?” relacionadas con la distribución de viajes.

El diagrama de la Figura 4.6 muestra de manera esquemática los componentes del simulador.

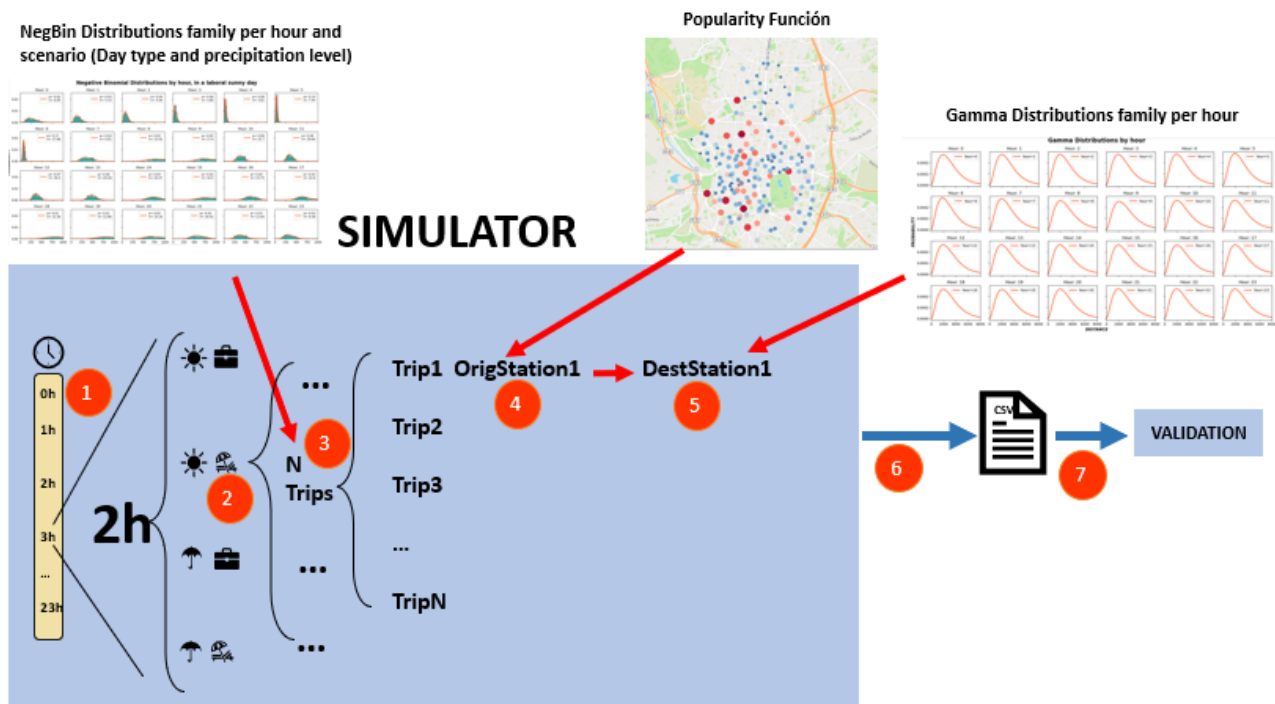


Figura 4.6: Diagrama que muestra los elementos constitutivos de nuestro simulador probabilístico: dado (1) la hora deseada; (2) las condiciones meteorológicas y laborales; (3) se estima el número (estocástico) de viajes; (4) se generan diferentes viajes según la popularidad de las estaciones hacia un (5) destino aleatorio. Finalmente, se (6) recopilan los datos y se (7) realizan algunas pruebas de validación. Fuente: Elaboración propia

A continuación, se detallan y explican los pasos del procedimiento:

- **Paso 1:** El procedimiento itera a través de cada hora del día.
- **Paso 2:** Para cada hora, se generan cuatro escenarios que combinan días lluviosos o no lluviosos con días laborales o festivos/fines de semana. Estos escenarios se utilizan para calcular los parámetros de las distribuciones binomiales negativas.
- **Paso 3:** Con la hora y el escenario actual, se consulta la Tabla 4.5 para obtener los valores de los parámetros p y n . Estos parámetros definen la distribución binomial negativa para esa hora y escenario, representando el número de viajes. Se genera un número entero aleatorio basado en dicha distribución, que representa el número de viajes en un día hipotético a la hora dada y bajo las condiciones proporcionadas. Al repetir este proceso múltiples veces, se obtienen diferentes números de viajes que, aunque siguen la misma distribución binomial, generan días contrafactuales correspondientes a situaciones hipotéticas.
- **Paso 4:** El identificador de la estación de origen se calcula para cada viaje generado en el paso anterior, basándose en una función de popularidad derivada de datos empíricos. La función de probabilidad

agrupa los datos según las estaciones de origen y la hora de los viajes. Para cada combinación de estación y hora, se calcula el conteo de viajes. Luego, se determina el porcentaje de popularidad de cada estación respecto al total de viajes registrados en todas las estaciones durante esa misma hora. Este proceso se repite de manera similar para las estaciones de destino (ver Apéndice I). Se sugiere que futuros estudios investiguen funciones de popularidad alternativas que integren variables socioeconómicas, geográficas, temporales y de comportamiento.

- **Paso 5:** Se selecciona la distribución Gamma asociada con la hora del viaje y se eligen únicamente las distancias relacionadas con los viajes cuyo origen corresponde al identificador de la estación determinado en el paso anterior mediante la función de popularidad. Se normalizan las probabilidades de este conjunto de distancias posibles (todas las combinaciones de viajes que comienzan en la estación de origen) y se genera un número aleatorio según esa distribución multinomial, obteniendo así una distancia de la cual se deriva la estación de destino. Como se explicó en el Capítulo 3, nuestra métrica de distancia caracteriza unívocamente cada viaje (distinguyendo la ida de la vuelta), por lo que la distribución multinomial devuelve automáticamente el destino de cada viaje.
- **Paso 6:** Al combinar todos los datos en un archivo CSV, se generan escenarios contrafactuales que corresponden a la distribución empírica de un año.
- **Paso 7:** Este paso incluye diversas validaciones: visuales y mediante el test de *Kullback-Leibler* de la distribución Gamma (Capítulos 4.2 y 4.2.1), la validación interna de la distribución NegBin (Capítulo 4.3.1) y la validación externa (Capítulo 5.2).

4.5. Descripción de alternativas para el ajuste de distribuciones con enfoque bayesiano

En la implementación del modelo, específicamente en el ajuste de distribuciones, se ha seguido un enfoque frecuentista utilizando la estimación por máxima verosimilitud (MLE) proporcionada por la función `fit()` de la biblioteca `scipy.stats`. En contraste, el enfoque Bayesiano incorpora información previa y actualiza el conocimiento a medida que se disponen de nuevos datos.

Esta sección explora la posibilidad de abordar el problema con un enfoque Bayesiano como alternativa al método utilizado en esta tesis. Es importante destacar el carácter exploratorio de este apartado, lo que implica que el modelo no se ha desarrollado con ambos enfoques. Su objetivo es discutir las diferencias entre ambos métodos y sentar las bases para considerar el enfoque Bayesiano como una alternativa viable para futuros trabajos.

Todo parte del Teorema de Bayes, una fórmula probabilística que se emplea para calcular la probabilidad condicional del evento A , dado que ocurre el evento B , expresada matemáticamente de la siguiente manera:

$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

Donde:

- $P(A | B)$ representa la probabilidad de que ocurra A dado que se ha observado o ha ocurrido B .
- $P(B | A)$ es la probabilidad de que ocurra B habiéndose dado A , conocida comúnmente como verosimilitud.

- $P(A)$ corresponde a la probabilidad previa o inicial de A antes de observar B .
- $P(B)$ constituye la probabilidad marginal de B independiente de cualquier otro evento. Actúa como un factor de normalización, garantizando que la suma de todas las probabilidades $P(A | B)$ sea igual a 1.

Al aplicar el Teorema de Bayes en la creación de modelos, $P(A | B)$ puede interpretarse como la probabilidad posterior de los parámetros (A) del modelo, habiéndose dado el conjunto de datos observados B .

Para concretar esta explicación y relacionarla con el tema central de la tesis, se describe a continuación un ejemplo práctico relacionado con el uso de bicicletas compartidas, aplicando los conceptos del Teorema de Bayes.

Por ejemplo, si el interés consiste en determinar la frecuencia media de uso de bicicletas (A) en BiciMad por hora durante un día, entonces:

- A es el parámetro del modelo (la frecuencia media de uso de bicicletas por hora en un día específico).
- B es el conjunto de datos observados (número de bicicletas utilizadas por hora en diferentes días).

A continuación se describe en la Tabla 4.7 cómo se interpreta y aborda el teorema de Bayes desde cada enfoque.

Concepto	Frecuentista	Bayesiano
Uso del teorema de Bayes	No se usa directamente en este enfoque. Se usa la verosimilitud para encontrar el valor que maximiza la probabilidad de observa B .	Lo primero es adaptar los elementos del Teorema de Bayes a terminología Bayesiana: $P(A B)$ se denomina “ <i>posterior</i> ” o probabilidad posterior, $P(B A)$ recibe el nombre de “ <i>likelihood</i> ” o verosimilitud, $P(A)$ se conoce como “ <i>prior</i> ” o probabilidad previa y finalmente $P(B)$ se conoce como “ <i>evidence</i> ” o evidencia. Este enfoque usa todo el Teorema de Bayes para, combinando el “ <i>prior</i> ” con la verosimilitud, obtener la distribución “ <i>posterior</i> ”.

Continúa en la siguiente página

Concepto	Frecuentista	Bayesiano
Método principal	Estimación de la Máxima Verosimilitud (<i>MLE</i> , por sus siglas en inglés, “ <i>Maximum Likelihood Estimation</i> ”). Matemáticamente se expresa como $\hat{A} = \arg \max_A P(B A)$, siendo $\hat{A}$ el estimador que maximiza la probabilidad $P(B A)$ e indicando $\arg \max_A$ que se busca el valor de A que maximiza $P(B A)$ (función de verosimilitud).	Cálculo de la distribución posterior $P(A B)$. Para ello se emplean algoritmos como “ <i>Monte Carlo Markov Chain</i> ” (<i>MCMC</i>), que aproximan la distribución posterior y simplifican su cálculo.
Evento A	Valor fijo discreto desconocido.	Variable aleatoria que incorpora incertidumbre. A se representa como una distribución de probabilidad.
Objetivo	Es estimar A basándose únicamente en la certeza de que se ha observado el evento B .	Es calcular la distribución posterior $P(A B)$
Probabilidad	Refleja la proporción de veces que el evento A ocurre en casos donde se ha observado B . Es decir, es la frecuencia relativa obtenida de las observaciones de las que se dispone.	Captura la incertidumbre asociada a A , no limitándose a una interpretación como frecuencia relativa. Por tanto es una medida subjetiva de creencia que puede actualizarse según se disponga de más información sobre B .
Información previa	No la utiliza ni la necesita.	Emplea una distribución previa $P(A)$ con el fin de estimar la posterior $P(A B)$.
Resultado	Valor discreto estimado $\hat{A}$ y su intervalo de confianza.	Distribución posterior $P(A B)$.
Cuándo se usa	Cuando se disponen de varias observaciones y no se tiene un conocimiento previo.	Cuando se cuenta con pocos casos observados pero existe conocimiento previo.

Continúa en la siguiente página

Concepto	Frecuentista	Bayesiano
Fortalezas y Debilidades	Como fortaleza de este enfoque se destaca la simplicidad, la eficiencia ante un gran volumen de datos y tener un coste computacional bajo. En contraposición tiene como debilidades que no se comporta correctamente si los datos son escasos, intervalos de confianza menos intuitivos y no permite fácilmente actualizar los resultados con nuevas evidencias.	Entre sus fortalezas destacan que permiten incluir información previa relevante al modelo, permite modelar la incertidumbre al ser el resultado una distribución y se adapta más fácilmente a nuevas evidencias. Por contrario, presenta debilidades como mayor complejidad, coste computacional más alto y una mala selección de los prior puede introducir sesgos.

Tabla 4.7: Comparativa del enfoque Frecuentista frente al enfoque Bayesiano para ajustar una distribución probabilística a unos datos empíricos.

Una vez contextualizados ambos enfoques se explora la posibilidad de obtener por ejemplo los parámetros de las distribuciones Gamma que modelan las probabilidades de viajar una cierta distancia empleando el servicio de BiciMad. En la sección 4.2.2 se obtuvieron dichos valores a nivel horario mediante el enfoque frecuentista. Es cierto que para obtener dichos valores se tuvieron en cuenta las horas a nivel anual (todo el dataset de 2018). Hay que recordar que el enfoque frecuentista necesita volumen de datos. Si se quisiera obtener distintas parametrizaciones a nivel de hora pero encuadradas en cada uno de los meses, el enfoque frecuentista empieza a ser menos robusto al tener menos observaciones (debilidad que ha sido descrita e la conceptualización de los enfoques en la Tabla tab:ComparativaFreq/Bay). Es por ello que se decidió explorar la posibilidad de obtener los parámetros de la distribución Gamma a nivel horario pero con una granularidad mayor que la anual, como puede ser la mensual o la semanal. El enfoque Bayesiano permitiría incluso obtener una distribución Gamma a eso niveles diferenciada por cada estación.

En el Apéndice J se aportan los códigos empleados para realizar dicha exploración. Como herramienta se ha utilizado el lenguaje de programación R concretamente con la interfaz RStan [The Comprehensive R Archive Network., 2024], que como su nombre indica, permite trabajar dentro de R con el lenguaje Stan [The Comprehensive R Archive Network., 2024] (plataforma de programación que permite y facilita la realización de inferencia estadística basada en modelos probabilísticos). Básicamente permite crear modelos con un lenguaje específico que luego es compilado y ejecutado por librerías C++. Estas características hacen que sea una herramienta ideal para la exploración que se desea realizar. Finalmente indicar que el nombre de este lenguaje es un homenaje a Stanislaw Ulam, un matemático polaco-estadounidense conocido por su contribución al desarrollo del método de Monte Carlo, realizada mientras trabajaba en el Laboratorio Nacional de Los Álamos en el contexto del Proyecto Manhattan [Wikimedia Foundation, Inc., 2024].

4.6. Resultados y conclusiones

Este capítulo se detalla la construcción del modelo probabilístico que predice la demanda del servicio de BiciMad. Tras analizar la literatura y los datos empíricos disponible se concluye que la demanda va a ser modelada empleando las variables de distancia recorrida en los viajes y número de viajes disponibles. La granularidad del modelo se establece en las distintas horas en punto, sin tener en cuenta los minutos, por una limitación de los datos disponible discutida en el Capítulo 3.2.1. Se decide realizar una inferencia de las distri-

buciones de probabilidad de cada una de esas variables con un enfoque frecuentista. Para lo cual se hace uso de algoritmo *MLE* disponible en el módulo `scipy.stats` de la biblioteca de Python `Scipy`. En el caso de la distancia, y tras evaluar las diferentes opciones de distribución empleadas en la literatura, se opta por emplear una distribución Gamma. Es la distribución que, tanto visualmente como cuantitativamente mediante la validación con el test de Kullback-Leibler, mejor se ajusta a los datos empíricos. Se justifica, además de por las validaciones realizadas, la idoneidad del uso de la Gamma fundamentando la decisión en el principio de máxima entropía. Seguidamente se modela el número de viajes que suceden en intervalos de una hora mediante una distribución binomial negativa (Negbin) considerando factores como días laborales, festivos y condiciones climáticas para obtener un ajuste más robusto. En este caso la idoneidad de esta distribución se verifica con una validación interna. Con la integración de ambas distribuciones y sumando información acerca de la popularidad a la hora de elegir una estación, se construye un simulador de viajes de BiciMad. Durante las validaciones externas a las cuales se somete el simulador, generando datos sintéticos para el año 2019 y comparándolos con los reales propios de dicho año, se describe una funcionalidad del mismo no prevista inicialmente. Dada su relevancia se decide describirla como una aplicación práctica del modelo descrita en el Capítulo X. Aunque se ha optado por un enfoque frecuentista por disponer de numerosos datos y ser una opción con menor coste computacional, también se realizan en este capítulo una exploraciones de ajustes de la distribución gamma con enfoque bayesiano. Tras comprobar que en el ordenador disponible este enfoque bayesiano, simplemente para realizar 7 agrupaciones a nivel de día de la semana invertía unas 40 horas de cómputo, y que el enfoque frecuentista arroja resultados robustos se decide mantener el simulador con el enfoque original. Quedan respondidas por tanto con el contenido de este capítulo las preguntas de investigación RQ4 y sus derivadas RQ4.1 y RQ4.2. formuladas en el Capítulo 1.5.2. También se da respuesta a los objetivos propuestos OB4.1 y OB4.2 descritos en el Capítulo 1.5.3.

“Any sufficiently advanced technology is indistinguishable from magic.”

Cualquier tecnología lo suficientemente avanzada es indistinguible de la magia.

— Arthur C. Clarke Tercera Ley de Clarke (Clarke’s third law)
Perfiles del futuro: Una indagación sobre los límites de lo posible
(*Profiles of the Future: An Inquiry into the Limits of the Possible*),
1973

5

Aplicaciones del Simulador - Modelo

Índice del capítulo

5.1. Simulación de la redistribución de viajes variando las estaciones disponibles	99
5.2. Imputación de datos y detección de atribuciones erróneas	103
5.3. Predicción de fallas en la demanda no satisfecha	104
5.4. Resultados y conclusiones	106

La verdadera utilidad de un modelo probabilístico no se limita exclusivamente a explicar las distribuciones de probabilidad que caracterizan distintos aspectos de un sistema complejo. Su potencial consiste en la capacidad de generar datos sintéticos que permitan explorar escenarios alternativos al utilizado como base para su propia creación.

En este capítulo se implementan aplicaciones prácticas del potencial del modelo probabilístico desarrollado, con el fin de constituir un aporte significativo a la gestión y optimización del servicio de BiciMad. Adicionalmente, las aplicaciones descritas se muestran como valiosas herramientas que pueden ayudar el problema de rebalanceo de bicicletas, clásico en los BSS.

Este capítulo aborda las siguientes preguntas: *Además de la predicción de la demanda de viajes, ¿Qué otras aplicaciones prácticas puede tener el modelo probabilístico que ayuden en la gestión y mejora del servicio BiciMad? ¿Qué utilidades prácticas puede tener la capacidad de generar escenarios contrafactuales (Qué pasaría si...)? ¿Cómo se puede emplear el modelo en el desarrollo de dichas aplicaciones y utilidades?*

5.1. Simulación de la redistribución de viajes variando las estaciones disponibles

En el contexto de la prestación del servicio de Bicimad, el cierre o la inoperatividad temporal de estaciones es un escenario posible, que además ya ha ocurrido en diversas ocasiones.

Existen diferentes razones y circunstancias que pueden explicar este cierre. Entre ellas se encuentran las intervenciones urbanísticas, como proyectos de asfaltado o reparación de vías, la realización de eventos variados, como competencias deportivas, desfiles o procesiones, las convocatorias de huelga, la falta de bicicletas en estación origen o de anclaje en la estación destino y los procesos de reubicación física de las estaciones.

En estos casos, los usuarios tienen la opción de utilizar un transporte alternativo o dirigirse a otra estación cercana, ya sea para recoger una bicicleta o para devolverla.

El problema radica en que el sistema de BiciMad (y en general cualquier sistema que involucre decisiones humanas) no permite registrar si el usuario ha optado por tomar o dejar una bicicleta en una estación diferente a la inicialmente prevista, debido a alguna de las circunstancias previamente mencionadas, entre otras posibles. Por tanto, también se desconoce a cuál estación alternativa con mayor probabilidad acudiría el usuario.

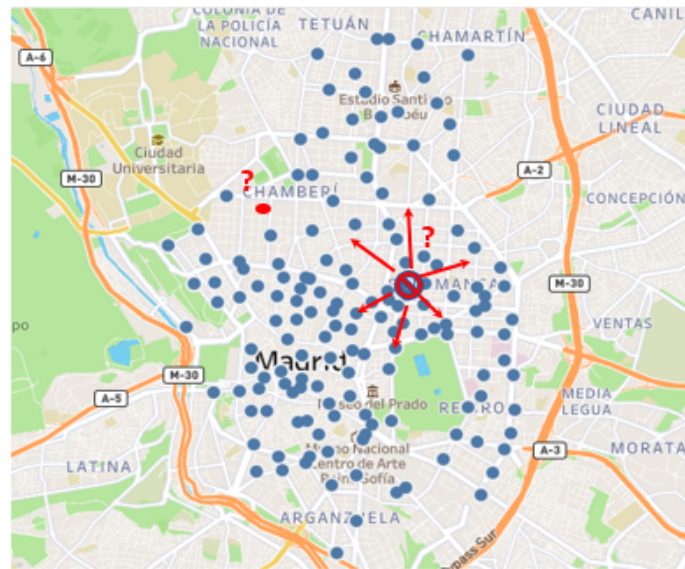
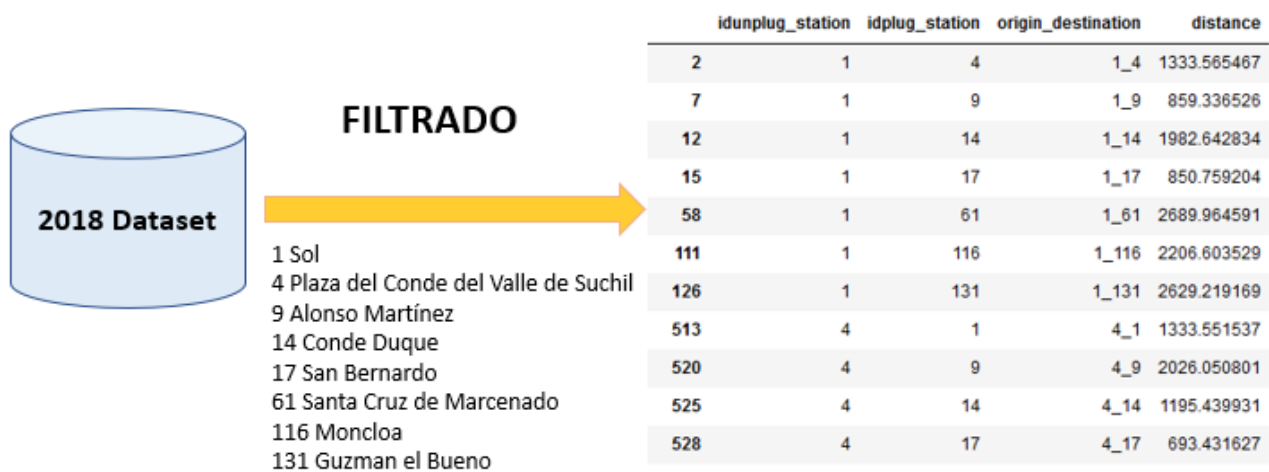


Figura 5.1: Visualización que ilustra la problemática de saber a qué estación acudiría el usuario ante la imposibilidad de acceder a inicialmente deseada, bien sea para tomar la bicicleta o para dejarla. Fuente: Elaboración propia.

Disponer de un simulador probabilístico capaz de generar escenarios contrafactuales permite simular *qué pasaría si* se elimina una estación. Por extensión, se puede recalcular la distribución probabilística de los viajes que involucraban dicha estación, entre otras estaciones activas. A continuación, se presenta un ejemplo simplificado para ilustrar el funcionamiento del simulador en un caso de este tipo. Dicho ejemplo consta de unos pasos que se describen a continuación:

1. **Paso 1:** Se seleccionó un subconjunto de estaciones a partir del conjunto completo de datos de 2018. La Figura 5.2 muestra el listado de estaciones seleccionadas y un resumen de las posibles combinaciones entre las estaciones seleccionadas y sus las distancias correspondientes entre ellas.



2018 Dataset → **FILTRADO**

1 Sol
4 Plaza del Conde del Valle de Suchil
9 Alonso Martínez
14 Conde Duque
17 San Bernardo
61 Santa Cruz de Marcenado
116 Moncloa
131 Guzman el Bueno

idunplug_station	idplug_station	origin_destination	distance
2	1	4	1_4 1333.565467
7	1	9	1_9 859.336526
12	1	14	1_14 1982.642834
15	1	17	1_17 850.759204
58	1	61	1_61 2689.964591
111	1	116	1_116 2206.603529
126	1	131	1_131 2629.219169
513	4	1	4_1 1333.551537
520	4	9	4_9 2026.050801
525	4	14	4_14 1195.439931
528	4	17	4_17 693.431627

Figura 5.2: Listado de estaciones seleccionadas y sus combinaciones con las distancias correspondientes. Fuente: Elaboración propia

2. **Paso 2:** Se determinó la probabilidad de viaje para cada par de estaciones utilizando la distribución Gamma, excluyendo aquellas en las que el origen y el destino son el mismo. Dado que la distribución Gamma es continua, se normalizó la probabilidad para reflejar únicamente las estaciones seleccionadas. La Figura 5.3 refleja la aplicación de este paso a ese subconjunto de registros que se empleó en la Figura 5.2 a modo representativo.

$$\Gamma_{pdf}(distancia)$$

$$\frac{\Gamma_{pdf}(distancia_i)}{\sum_{n=0}^N \Gamma_{pdf}(distancia_n)}$$

Normalizado

idunplug_station	idplug_station	origin_destination	distance	Prob	ProbN1
2	1	4	1_4 1333.565467	0.000218	0.018739
7	1	9	1_9 859.336526	0.000146	0.012597
12	1	14	1_14 1982.642834	0.000260	0.022393
15	1	17	1_17 850.759204	0.000145	0.012465
111	1	116	1_116 2206.603529	0.000261	0.022454
126	1	131	1_131 2629.219169	0.000248	0.021348
513	4	1	4_1 1333.551537	0.000218	0.018739
520	4	9	4_9 2026.050801	0.000261	0.022446
525	4	14	4_14 1195.439931	0.000200	0.017225
528	4	17	4_17 693.431627	0.000116	0.009974

Figura 5.3: Probabilidad de viaje entre pares de estaciones según la distribución Gamma, normalizada para considerar solo las estaciones seleccionadas. Fuente: Elaboración propia

3. **Paso 3:** Posteriormente, se recalculó la probabilidad normalizada de la misma manera, pero eliminando la estación de Santa Cruz de Marcenado (ID 61). Esta exclusión tiene como objetivo simular una

situación en la que un usuario inicia un viaje en la estación de Sol (ID 1) con destino a Santa Cruz de Marcenado, pero al llegar no puede estacionar la bicicleta debido a la falta de anclajes libres disponibles, viéndose obligado a dirigirse a otra estación. La Figura 5.4 refleja el resultado de dicho cálculo.

Normalización pero
quitando la estación 61
" Santa Cruz de Marcenado "

Diferencia entre
Normalización con y sin
la estación eliminada

idunplug_station	idplug_station	origin_destination	distance	Prob	ProbN1	ProbN2	ProbN2-ProbN1	
0	1	116	1_116	2206.603529	0.000261	0.022454	0.029485	0.007032
1	1	14	1_14	1982.642834	0.000260	0.022393	0.029405	0.007012
2	1	131	1_131	2629.219169	0.000248	0.021348	0.028033	0.006685
3	1	4	1_4	1333.565467	0.000218	0.018739	0.024607	0.005868
4	1	9	1_9	859.336526	0.000146	0.012597	0.016541	0.003945
5	1	17	1_17	850.759204	0.000145	0.012465	0.016368	0.003903
6	1	61	1_61	2689.964591	0.000245	0.021088	0.000000	-0.021088

Figura 5.4: Recálculo de la probabilidad normalizada eliminado la estación de Santa Cruz de Marcenado (ID61) y diferencia con respecto a la normalización que la incluye. Obsérvese que, en el registro 6 que corresponde a las estaciones de Origen Sol (ID1) y destino Santa Cruz de Marcenado (ID61), se ha puesto la probabilidad a valor cero por no poder darse esta situación, lo que hace que la diferencia sea negativa. Fuente: Elaboración propia

4. **Paso 4:** Finalmente se observa en la Figura 5.5 cómo, de acuerdo con la distribución Gamma, los destinos de los viajes cuyo origen es la estación de Sol se redistribuyen entre las diferentes estaciones. Debido a que la estación de Santa Cruz de Marcenado no está disponible, los viajes se redistribuyen entre las estaciones activas, teniendo mayor coincidencia aquellas que están más cercanas al destino deseado.

Mediante la aplicación del simulador probabilístico, se ha logrado modelar y analizar con detalle la redistribución de los viajes cuando una estación deja de estar operativa en el sistema de BiciMad. Este análisis permite evaluar cómo varían las probabilidades de destino al eliminar una estación específica, en este caso, Santa Cruz de Marcenado. También permite mostrar en un mapa cómo dichas probabilidades se redistribuyen hacia otras estaciones activas. Los resultados obtenidos destacan que las estaciones más cercanas al destino inicial suelen concentrar la mayoría de los viajes. Este hecho concuerda con lo que se consideraría un comportamiento lógico y predecible por parte de los usuarios.

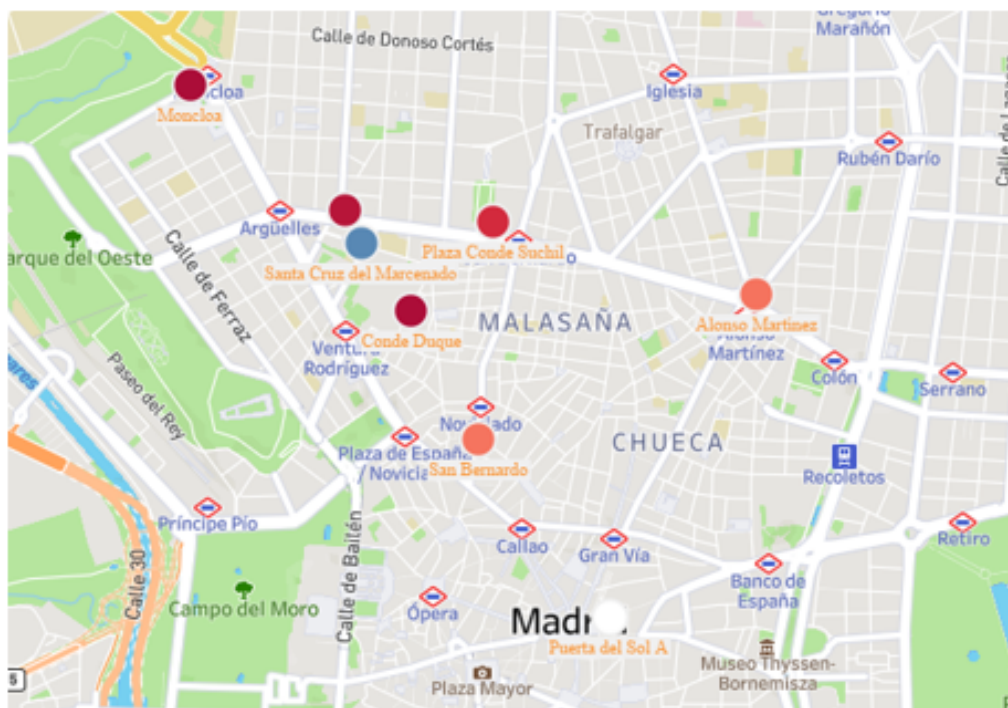


Figura 5.5: Visualización de la redistribución de probabilidades entre las estaciones activas más cercanas a la estación eliminada. La estación de Santa Cruz de Marcenado, eliminada en la simulación, se destaca en color azul. Los puntos en tonos naranja-rojo representan las posibles estaciones destino para los viajes originados en Sol, al no poder finalizar en la estación eliminada. La intensidad del color indica la probabilidad, siendo más oscura cuanto mayor es ésta. Fuente: Elaboración propia

5.2. Imputación de datos y detección de atribuciones erróneas

Una de las principales ventajas de los modelos probabilísticos, como el desarrollado en el Capítulo 4, es su capacidad para generar datos sintéticos. Estos datos, aunque simulados, mantienen coherencia con las distribuciones probabilísticas ajustadas en el modelo, lo que permite crear escenarios contrafactuales robustos respaldados por dichas distribuciones. Por lo tanto, estos modelos no solo facilitan la comprensión de sistemas complejos, sino que también ofrecen una solución a los problemas de datos faltantes y a la necesidad de imputarlos, muy comunes en el análisis de datos.

Tradicionalmente, se emplean técnicas como el cálculo del promedio o la mediana de los datos disponibles, la regresión lineal o logística, y los métodos basados en árboles de decisión para imputar datos faltantes. Sin embargo, recurrir a un modelo probabilístico para resolver este problema asegura la coherencia interna de los datos y refuerza el realismo y la relevancia de los resultados obtenidos. Cabe destacar que los datos generados con modelos probabilísticos podrían haberse dado en un escenario alternativo o paralelo, es decir, podrían haber sido perfectamente reales.

En el caso particular de BiciMad, el modelo podría utilizarse, por ejemplo, para imputar datos faltantes durante el período de suspensión del servicio debido a incidencias y confinamientos por COVID-19. Se generarían viajes sintéticos durante ese período, considerando la tipología de días y las condiciones climáticas, alineados con las distribuciones de probabilidad previas al confinamiento. Asignando un coste o precio promedio a cada viaje generado, sería posible estimar el dinero dejado de ingresar durante ese periodo. Además, otra línea de análisis interesante sería comparar los datos reales de viajes realizados durante el período de

reapertura y vuelta a la normalidad con los generados por el simulador. Esto permitiría medir, por ejemplo, el tiempo que el servicio necesitó para alcanzar los niveles de operación previos a la pandemia.

Aunque en los párrafos anteriores se presenta un caso específico de imputación relacionado con BiciMad, se decidió no implementarlo debido a que se encontró, de forma inesperada, un ejemplo más interesante y significativo, que se describe a continuación.

Después de llevar a cabo la validación interna del modelo probabilístico en el Capítulo 4.3.1, se realizaron pruebas de validación externa con datos nuevos de 2019. Debido a que en 2019 se agregaron nuevas estaciones al servicio de BiciMad, según se refleja en el Apéndice B, se filtraron los datos de 2019 para mantener únicamente aquellos viajes que tuvieran como origen y destino una de las estaciones disponibles en 2018. Se conservaron las estaciones con ID comprendidos entre el 1 y el 175, excepto las estaciones 22, 68 y 70 que se eliminaron en 2019. Por lo tanto, las estaciones con identificador entre 176 y 219 no se consideraron en esta validación al haberse introducido por primera vez en 2019.

El proceso de validación comenzó comparando la distribución del número de viajes a lo largo de 2018, agrupados por hora del día y representados mediante un diagrama de cajas (ver diagrama de cajas superior de la Figura 5.6). Este tipo de diagramas resalta variaciones, medianas y posibles valores atípicos en cada hora. Al comparar esta distribución con el diagrama de cajas correspondiente de los datos filtrados de 2019 (ver diagrama de cajas central de la Figura 5.6), se apreciaba un cambio significativo en el número de viajes entre las 6 y las 8 de la mañana. No obstante, desde una perspectiva habitual, este incremento resulta poco coherente dada la franja horaria, que normalmente presenta una demanda de viajes relativamente baja. Este hecho motivó un análisis exhaustivo tanto a nivel de viajes como de horas, identificando un total de 107 días anómalos en 2019. La mayor concentración de estas anomalías se produjo entre el 27 de noviembre y el 20 de diciembre. La exploración detallada de los viajes realizados en los días considerados anómalos llevó a la hipótesis de que un ajuste de +2 horas en esos días específicos podría restablecer la coherencia entre los datos de 2018 y 2019. Por lo tanto, se aplicó dicho desfase horario, se reconstruyeron las series temporales y se generó nuevamente la gráfica para 2019 (ver diagrama de cajas inferior de la Figura 5.6). Como resultado, los valores de 2019 ahora se encuentran dentro del rango esperado en comparación con los de 2018.

Esta observación se confirmó posteriormente cuando nos comunicamos con BiciMad para consultar la posibilidad de que hubiera datos erróneos en 2019. El administrador de datos de BiciMad nos remitió un mensaje publicado en la plataforma el 9 de septiembre de 2021, el cual corroboraba nuestros hallazgos. En la Figura 5.7 puede verse una captura de dicho comunicado.

Como se mencionó anteriormente, se consideró más adecuado incluir este ejemplo en este capítulo en lugar de abordar la imputación de datos descrita previamente, ya que representa una aplicación más avanzada de nuestro modelo.

Como resultado de nuestros hallazgos y de la investigación posterior que se han descrito en este capítulo y en nuestro segundo artículo enmarcado en esta tesis [Vallez et al., 2024], se identificó una aplicación práctica de nuestro modelo que no se había previsto inicialmente. En concreto, el modelo demuestra su capacidad para validar datos del mundo real y sugiere la posibilidad de que existan datos erróneos cuando los valores observados se salen de los rangos establecidos de forma probabilística. Esta capacidad de validación amplía el uso del modelo más allá de la imputación de datos.

5.3. Predicción de fallas en la demanda no satisfecha

En esta aplicación práctica final, se investigó la posibilidad de que se produzca una demanda no satisfecha debido a la falta de bicicletas disponibles en las estaciones seleccionadas para iniciar un viaje o a la insuficiencia de puntos de anclaje para dejar las bicicletas en las estaciones elegidas para finalizar el trayecto.

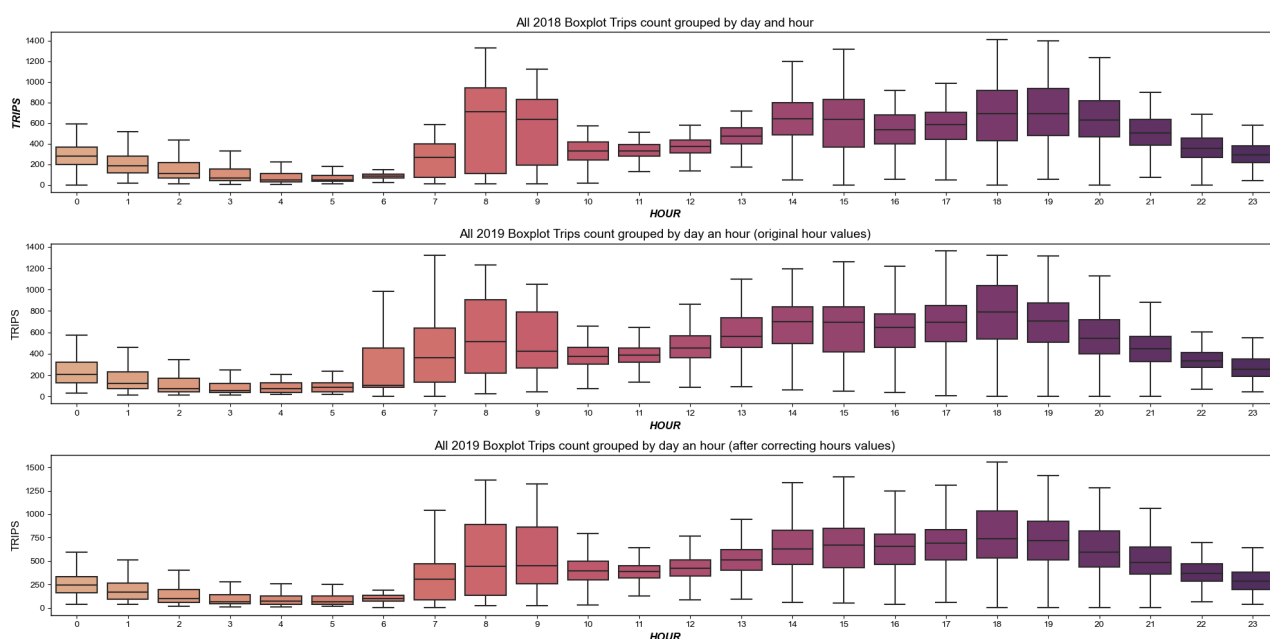


Figura 5.6: Distribución del número de viajes por hora en 2018 (superior), 2019 (central) y 2019 después de corregir datos mal atribuidos (inferior). La comparación KL de los diagramas de cajas en el rango de 6 a 8 a.m. ayudó a identificar grandes cambios en esas horas. Después de contactar al proveedor de datos, se confirmó que parte de los datos (después del horario de verano de octubre) estaban etiquetados incorrectamente. En la figura inferior, se volvió a graficar los datos tras esta corrección, y la distribución coincide estrechamente con la de 2018, de la misma manera que la distribución de distancias en 2018 y 2019 es casi idéntica (ver Fig. 3.1). Fuente: Elaboración propia a partir de los datos de viajes de BiciMad de los años 2018 y 2019.

MADRID | Portal de datos abiertos del Ayuntamiento de Madrid

¿Qué estás buscando?

En portada
 Acerca de Datos Abiertos
 Catálogo de datos
Colabora
 Visualiza
 Solicitud de reutilización

Catálogo de datos > Conjuntos de datos

Administrador Datos Abiertos 09/09/2021 11:36:10 [permalink](#)

A veces en el sistema BiciMAD se aplican regularizaciones como resultado de acciones relacionadas con el servicio de atención al cliente, por este motivo, si se realiza un ajuste en el instante de anclaje o desanclaje, podrían aparecer este tipo de efectos. También puede ocurrir durante el cambio horario estacional, cuando se aplica el retraso de hora. Gracias por su interés y colaboración en la mejora del servicio.

Figura 5.7: Captura del mensaje publicado en la plataforma de datos de BiciMad el 9 de septiembre de 2021 corroborando la posibilidad de datos erróneos. Fuente: Portal de datos abiertos del Ayuntamiento de Madrid

Se presenta a continuación los pasos realizados para el desarrollo de una prueba de concepto cuyo objetivo es demostrar la viabilidad y aplicabilidad del modelo en la predicción de fallas en la demanda no satisfecha.

1. **Paso 1:** Calcular del número promedio de bicicletas disponibles y de puntos de anclaje vacíos por hora para cada estación: Este cálculo se lleva a cabo realizando numerosas solicitudes, a intervalos de 5 minutos durante varios días, a la API descrita en el Capítulo 3.2.5, mediante un programa desarrollado específicamente para este propósito. Hay destacar que, tal como se indicó en el capítulo referenciado, dicho programa proporciona archivos JSON con información actualizada sobre las estaciones en intervalos de 5 minutos. Dado que el modelo desarrollado en esta tesis, descrito en el Capítulo 4, opera a nivel horario y no considera los minutos exactos en los que ocurre cada viaje, debido a una limitación en los datos explicada en el Capítulo 3.2.1, es necesario calcular los promedios de los valores proporcionados por el JSON cada 5 minutos para bicicletas disponibles y puntos de anclaje vacíos, obteniendo así un valor representativo por hora.
2. **Paso 2:** Recopilar datos meteorológicos de esos días y los categorizamos en fines de semana/días festivos y días laborables.
3. **Paso 3:** Emplear el simulador, para generar varios escenarios contrafactuales de viajes para esos días, incorporando sus características reales.
4. **Paso 4:** Finalmente, comparar el volumen de viajes en la estación de origen y la hora correspondiente en los escenarios sintéticos con la disponibilidad real de bicicletas en la estación de origen. De manera similar, se compara el volumen de viajes en la estación de destino y la hora en los escenarios sintéticos con la disponibilidad real de puntos de anclaje vacíos en la estación de destino.

Los hallazgos indican que, en general, la demanda se satisface a pesar de la variabilidad introducida por la naturaleza probabilística de los datos generados. En casos extremos, la demanda no se cumplió por apenas 1 a 3 bicicletas o puntos de anclaje. Esto sugiere que el servicio ha alcanzado un nivel de autorregulación, aunque también podría existir una oportunidad para estrategias más orientadas al riesgo que aumenten el número de usuarios, aprovechando la alta calidad del servicio ofrecido por BiciMad.

Dado que esta es una prueba de concepto, los datos carecen de significancia estadística suficiente para ser considerados completamente relevantes. Para futuros trabajos, se recomienda ampliar esta idea realizando simulaciones a mayor escala con una significancia estadística más robusta. Sin embargo, esta prueba de concepto establece una metodología que puede ayudar a identificar la demanda insatisfecha del servicio, tanto en bicicletas disponibles en la estación de origen como en puntos de anclaje vacíos en la estación de destino.

5.4. Resultados y conclusiones

Tras la implementación del modelo probabilístico y su posterior inclusión en el simulador de viajes de BiciMad detallado en el Capítulo 4.4, resultaba necesario comentar y desarrollar aplicaciones prácticas para el mismo. No era suficiente con limitar la funcionalidad práctica al ámbito de explicar un sistema complejo como es el servicio ofrecido por BiciMad. Por tanto, se tomó la capacidad principal del mismo que es la generación de datos sintéticos y se exploró como estos datos podían ser utilizados de forma práctica por parte del gestor del servicio de BiciMad. La primera aplicación práctica permite a BiciMad, simular cómo eventos que afectan a la disponibilidad del servicio en estaciones afecta a la demanda redistribuyéndola en las estaciones más cercanas. Es en definitiva una herramienta con potencial que permite anticiparse al comportamiento de demandas futuras que se generan en escenarios contrafactuales. Y todo ello sin incurrir en

gastos de análisis de encuestas o recopilación de datos a pie de estación, obteniendo resultados robustos basados en probabilidad y que incorporan la incertidumbre inherente al propio sistema. La segunda aplicación práctica está relacionada con la capacidad que tiene el modelo de imputar datos. Lo único que el caso de uso elegido para demostrar la realización de la imputación deriva de un descubrimiento obtenido al realizar la validación externa del modelo frente a los datos de 2019. este descubrimiento consiste en la capacidad que tienen los datos sintéticos de apuntar posible anomalías o desviaciones de lo esperado en los datos reales. Se consiguió identificar que ciertos viajes de 2019 tenían una hora de viaje mal asignada, hecho que fue validado en la plataforma de datos abiertos de BiciMad, por el administrador de la misma. Finalmente y para la tercera aplicación práctica se decide enunciar una metodología, validada con la implementación de una prueba de concepto, para cuantificar la demanda que no ha sido cubierta. tanto si se debe a falta de bicicletas en la estación origen como si es motivada por la falta de anclajes en la de destino. estas aplicaciones dan respuesta a la pregunta de investigación RQ5 enunciada en el Capítulo 1.5.2 y permiten completar los objetivos identificados comprendidos entre el OB5.1 y el OB5.4 incluidos, descritos en el Capítulo 1.5.3.

“oída ouk eidos”.

Solo sé que no sé nada.

— Sócrates *Apología de Sócrates (Apología Sokratous)*, recogido
por Platón 399 a.C

6

Conclusiones, contribuciones y futuros desarrollos

Índice del capítulo

6.1. Conclusiones de la tesis	108
6.2. Principales contribuciones de esta tesis	110
6.3. Publicaciones y conferencias	111
6.4. Futuras líneas de investigación	112

6.1. Conclusiones de la tesis

En esta sección se discuten primeramente las conclusiones a nivel de capítulo, terminando con las conclusiones globales a nivel de investigación.

Del Capítulo 1 “Introducción y motivación” se evidencia que el surgimiento y posterior éxito de los BSS se debe gran parte a la concienciación social de la necesidad de ser sostenibles en todos los aspectos incluso en el transporte. Aunque los BSS se presentan como la implementación mas exitosa del concepto de “*vehicle-sharing*” a nivel global, queda constatado de que existen diversas tipologías de sistemas y un proceso de evolución a lo largo de los años. Al ser abordados muchos estudios sobre BSS por parte de varios autores con herramientas de aprendizaje automático, era obligado y necesario disponer de una mínima introducción y clasificación de dichas herramientas.

Respecto al Capítulo 2 “Estado del arte” Los estudios acerca de los BSS tiene la característica de poder abarcar diversos ámbitos de conocimiento, siendo el problema del rebalanceo el principal desafío de este ámbito de estudio. Una conclusión directa que se extrae de esta parte es que un problema puede tener múltiples soluciones, cada una con sus fortalezas y debilidades. De ahí la importancia de tener unas buenas revisiones de la literatura que permitan conocer lo realizado y con qué alcance. Si se dispone, como ha sido el caso, de una clasificación de las herramientas posibles a utilizar (algoritmos de aprendizaje automático) y un análisis de los estudios realizados hasta ahora (estado del arte), se facilita la posibilidad de identificar investigaciones pendientes de realizar. En este caso esa investigación pendiente se ha traducido en el uso de métodos

probabilísticos en vez de los deterministas ampliamente usados.

En lo referente al Capítulo 3 “Análisis exploratorio de los datos (EDA)” destacar como conclusiones la importancia de la disponibilidad de los datos para poder implementar un modelo probabilístico. En esta línea surgió la necesidad de disponer una variable que representase la distancia. Aun habiendo diversos métodos para obtener distancias entre dos puntos dada su longitud y latitud, fue necesario implementar un cálculo que permitiera discriminar y diferenciar la distancia entre un punto A y otro B en función del sentido del viaje.

Del Capítulo 4 “Enfoque de ML probabilístico para el modelado de la demanda de BiciMad” seguramente sea del que más conclusiones se han podido extraer. primeramente los dos enfoques existentes para inferir distribuciones de probabilidad cada uno con sus fortalezas y limitaciones. Adicionalmente se pudo constatar que en el caso de BiciMad es la distribución Gamma la que mejor se ajusta a los datos de distancias. Sin embargo existen artículos en la literatura que demuestran que en otras implementaciones otras distribuciones realizan un mejor ajuste. Esto lleva a pensar que existe alguna característica propia de la implementación de BSS estudiada que le confiere la cualidad de acercarse más a una u otra distribución. También se constata que la justificación de la distribución Gamma en este caso se tiene que apoyar en el principio de máxima entropía, no siendo posible aplicar la interpretación de la Gamma como suma de varias exponenciales. Con respecto a los viajes la Binomial Negativa es la que aporta un mejor ajuste siempre y cuando se discrimine por tipo de día y si ha habido o no lluvia. Todo ello permite crear un modelo que unido a una función de popularidad de las estaciones resulta en un simulador de viajes de BiciMad. Aunque se han explorado algunas alternativas no se ha encontrado una función de Popularidad mejor que la resultante de aplicar la frecuencia de aparición en los distintos viajes. No obstante este es un claro aspecto a ser tratado como trabajo futuro. Para finalizar con las conclusiones obtenidas en este capítulo se ha podido verificar el coste computacional que supone el enfoque bayesiano para el ajuste de distribuciones, si bien compensa al aportar una granularidad mayor por no tener una dependencia del número de observaciones que se tienen de casos particulares. este punto también se transforma directamente en un trabajo futuro.

Finalmente y en relación al Capítulo 5 “Aplicaciones del Simulador - Modelo” la conclusión primordial que se extrae es la versatilidad que tienen los modelos probabilísticos para contribuir a aplicaciones prácticas más allá de la simple generación de datos sintéticos. Refuerza sin duda la idea de que la naturaleza obedece de alguna forma a patrones definidos por distribuciones estadísticas y que todo puede ser visto y analizado desde la óptica de la probabilidad.

En definitiva esta tesis, y de forma general, aporta las siguientes conclusiones:

- El campo de conocimiento relacionado con los BSS es amplio y heterogéneo. Lo cual lo convierte en atractivo para realizar futuras investigaciones y continuar con la línea iniciada en esta tesis.
- Los BSS tienen una gran influencia de la localización donde se implementan. Lo que funciona en una localización puede resultar estéril en otra.
- Los viajes que se producen en los BSS son realizados por personas con características heterogéneas. Se pueden emplear un enfoque individualista tratando de caracterizar el viaje de cada individuo procesando varias variables o se puede emplear un enfoque poblacional basado en distribuciones de probabilidad e incorporando la incertidumbre en el análisis.
- Modelos probabilísticos que generen datos sintéticos puede ser de gran ayuda para mejorar la eficacia de los procesos de optimización del servicio.
- La analítica de datos es una poderosa herramienta que tiene aportes en prácticamente cualquier disciplina.

- El modelo propuesto demuestra varias fortalezas clave que lo convierten en un enfoque valioso para analizar matrices Origen-Destino (O-D). En primer lugar, tiene la capacidad de generar datos sintéticos con una mayor explicabilidad y precisión, lo que mejora la comprensión de los patrones de viaje entre orígenes y destinos. Además, el modelo integra la distribución de datos dentro de su estructura, permitiéndole capturar de manera efectiva la aleatoriedad inherente a los datos de viajes. Asimismo, mitiga los desafíos clásicos asociados con los modelos de gravedad, como la complejidad introducida por múltiples variables de entrada y la dificultad de explicar diversos escenarios utilizando información de viajes a nivel individual. Esto hace que el modelo propuesto sea tanto versátil como robusto en su aplicación.
- Sin embargo, el modelo también presenta algunas limitaciones que requerirán mayor atención. Su escalabilidad representa un desafío, ya que manejar conjuntos de datos más grandes demandaría recursos computacionales significativos. Además, el modelo requiere un manejo cuidadoso para evitar el sobreajuste y reducir el riesgo de simplificar en exceso elementos superpuestos dentro de la distribución de datos. Abordar estas limitaciones será crucial para mejorar la aplicabilidad del modelo y garantizar su efectividad en otros contextos, como, por ejemplo, en sistemas de bicicletas compartidas sin estaciones, donde las bicicletas pueden ser tomadas o dejadas en cualquier lugar.

6.2. Principales contribuciones de esta tesis

A continuación, se presenta un listado con las principales contribuciones realizadas en esta tesis:

- Crear un tesoro para unificar la terminología relacionada con los BSS, ampliando así el alcance de la revisión del estado del arte al abordar la fragmentación terminológica identificada en este campo.
- Desarrollar una metodología para la revisión sistemática de la literatura que utilice herramientas de visualización y agrupación cronológica y temática de los artículos.
- Confirmar que el problema del rebalanceo es el desafío central en el ámbito de los BSS y verificar, mediante la revisión de la literatura, que este problema se aborda abordando cada uno de sus subproblemas: la predicción de la demanda del servicio y la optimización del reparto de bicicletas en las estaciones para satisfacer la demanda minimizando los costes asociados.
- Elaborar una taxonomía de los algoritmos propuestos en la literatura para abordar el problema del rebalanceo, destacando la preferencia por algoritmos deterministas sobre los probabilísticos.
- Presentar un enfoque simplificado que fundamenta la explicación de la demanda en un servicio de BSS como BiciMad, considerando las variables distancia recorrida y número de viajes por unidad temporal, en contraste con los enfoques clásicos que analizan múltiples variables para capturar la heterogeneidad de estos servicios.
- Implementar un método de cálculo de distancias que genera valores distintos según la dirección del trayecto (de A a B o viceversa), utilizando la longitud y latitud de los puntos origen y destino, basado en las distancias Manhattan y Haversine.
- Desarrollar un marco de aprendizaje automático probabilístico para analizar y simular la demanda en BSS, justificando el uso de las distribuciones Gamma y Binomial negativa para modelar la distancia y el número de viajes, respectivamente. Este modelo mitiga los desafíos de los modelos gravitacionales

clásicos, como la complejidad de múltiples variables de entrada y la dificultad para explicar diversos escenarios con datos de viajes individuales.

- Calcular y validar los parámetros de las distribuciones probabilísticas ajustadas a los datos empíricos de BiciMad del año 2018.
- Desarrollar un simulador de viajes de BiciMad integrando el modelo probabilístico presentado con información sobre la popularidad de las estaciones, permitiendo predecir escenarios hipotéticos y realizar análisis contrafactuales que faciliten la toma de decisiones en la planificación del servicio.
- Aplicar el simulador creado en diversas situaciones prácticas:
 - Simular la redistribución de viajes variando las estaciones disponibles.
 - Imputar datos faltantes.
 - Detectar atribuciones erróneas en los datos.
 - Predecir fallas en la demanda no satisfecha.

6.3. Publicaciones y conferencias

Las siguientes publicaciones y presentaciones en conferencias han sido resultado de la investigación desarrollada en el marco de esta tesis doctoral.

1. *Challenges and opportunities in dock-based bike-sharing rebalancing: a systematic review (2021)*

Autores: Vallez-Fernández, Carlos Miguel; Castro-Ponce, Mario; Contreras-Bárcena, David

Sustainability (MDPI), 2021, vol. 13, no 4, p. 1829. Colección: *Bicycling Today: Establishing and Learning from the State of the Art*

JCR Impact Factor 2021: 3.2

5 Years JCR Impact Factor: 3.6

JCR category rank: Q2

Scopus SJR: 0.664

Scopus CitiScore 2021: 5.0

H5-Index: 223

DOI: <https://doi.org/10.3390/su13041829>

El artículo ha sido citado: 35 veces

2. *A Probabilistic Machine Learning Framework for Modeling Bicycle Sharing Demand: The BiciMad Case Study (2025 en revisión)*

Autores: Vallez-Fernández, Carlos Miguel; Castro-Ponce, Mario; Contreras-Bárcena, David

Computers & Industrial Engineering (*artículo en revisión*)

JCR Impact Factor 2023: 6.7

JCR category rank: Q1

Scopus SJR: 1.76

Scopus CitiScore 2023: 12.7

H5-Index: 148

DOI: (*artículo en revisión*)

3. *El principio de máxima entropía y la distribución de viajes en servicios de bicicletas compartidas*

Formato: Póster

Autores: Vallez-Fernández, Carlos Miguel; Castro-Ponce, Mario; Contreras-Bárcena, David

XXIII Congreso de Física Estadística - FisEs 2022

Organizado por: Universidad de Zaragoza Desde 12/05/2022 hasta 14/05/2022. .

Además de las publicaciones y el congreso previamente mencionados y directamente relacionados con esta tesis, he participado en otras publicaciones académicas que no están enmarcadas en el trabajo presentado en esta tesis. Se enumeran a continuación:

1. *Unveiling the effect of social media communication on urban mobility (2024)*

Autores: Martínez de Ibarreta-Zorita Carlos; Cifuentes-Quintero, Jenny Alexandra; Vallez-Fernández, Carlos Miguel; Betancourt-Odio, Manuel Alejandro

Transportation (Springer), 2024, p. 1-45.

JCR Impact Factor 2023: 3.5

5 Years JCR Impact Factor: 4.3

JCR category rank: Q1

Scopus SJR: 1.475

Scopus CitiScore 2023: 10.7

H5-Index: 105

DOI: <https://doi.org/10.1007/s11116-024-10512-6>

2. *Examen 3.0: Aplicación para la generación automática de exámenes complejos en plataforma Moodle*

Formato: Virtual

5º Congreso Virtual Internacional de Educación, Innovación y TIC, EDUNOVATIC2020

organizado por: REDINE, Red de Investigación e Innovación Educativa, con la colaboración de la editorial Adaya Press Desde 10/11/2020 hasta 11/11/2020.

Contribución recogida en un Libro de Proceedings con ISBN 978-84-09-22967-3.

6.4. Futuras líneas de investigación

Considerando que esta tesis integra dos artículos diferenciados pero a la vez complementarios dentro del ámbito de los BSS, las propuestas para trabajos futuros se estructuran en dos grupos:

- 1. Revisión sistemática y estudios de los BSS relacionado con el primer artículo “*Challenges and Opportunities in Dock-Based Bike-Sharing*” [Vallez et al., 2021].
- 2. Modelización probabilística, el problema del rebalanceo y la implementación de BiciMad relacionados con el segundo estudio “*Probabilistic Machine Learning Framework for Modeling Bicycle Sharing Demand*” [Vallez et al., 2024].

A continuación se detallan las propuestas para ser abordadas en futuras líneas de investigación agrupadas en las dos categorías anteriormente comentadas:

1. Relacionadas con la revisión sistemática y los estudios de los BSS.

- 1.1 Resulta llamativo que los trabajos existentes abordan los BSS como una alternativa sostenible. Sin embargo muchas implementaciones emplean vehículos contaminantes para realizar el proceso de reequilibrio de bicicletas entre las diferentes estaciones. Sería interesante estudiar dicho impacto y compararlo con la reducciones de emisiones contaminantes derivadas del uso de este tipo de servicios.
- 1.2 Desarrollar estrategias híbridas de reequilibrio combinando enfoques temporales estáticos y dinámicos por un lado y adicionalmente combinando también enfoques de responsabilidad del equilibrio (realizados por la empresa suministradora del servicio o por los usuarios mediante incentivos).
- 1.3 Estudiar el impacto de diferentes incentivos económicos para usuarios en el reequilibrio basado en demanda. Creación de simuladores que permitan ayudar a los proveedores del servicio a tomar decisiones sobre el sistema tarifario implementado.
- 1.4 El "problema del rebalanceo" se ha estudiado en otros ámbitos como por ejemplo en otros sistemas de vehículos compartidos y en problemas clásicos de recogida y entrega en servicios de mensajería. Se propone la realización de estudios comparativos que analicen las soluciones adoptadas en otros campos y cómo podrían aplicarse a los BSS.
- 1.5 Realización de revisiones y estudios interdisciplinarios, por ejemplo salud y economía u optimización y sostenibilidad por ejemplo.
- 1.6 Profundizar en estudios comparativos entre diferentes implementaciones de BSS. El objetivo es establecer una taxonomía según parámetros propios de la implementación como el número de estaciones, y la extensión geográfica entre otros factores. Este enfoque busca identificar patrones comunes y por ejemplo ser capaces de definir qué se considera un BSS grande o pequeño. esto permitiría analizar los diferentes comportamientos asociados y, además, inferir características específicas en función del subgrupo de BSS al cual pertenezcan.
- 1.7 Creación de una infraestructura digital de datos para facilitar el intercambio de información en diferentes sectores y ámbitos de investigación

2. Modelización probabilística, el problema del rebalanceo y la implementación de BiciMad.

- 2.1 Extender el modelo probabilístico a sistemas de BSS en diversas ciudades y evaluar su generalizabilidad.
- 2.2 El artículo [[Kou and Cai, 2019](#)] analiza las distribuciones de distancias en diversas ciudades y observa que, según el tamaño de la implementación del BSS, algunas distribuciones se ajustan mejor que otras. Como línea de trabajo futura, se propone desarrollar una meta distribución que, considerando parámetros específicos del BSS como su tamaño, se adapte flexiblemente a diferentes implementaciones.
- 2.3 Emplear herramientas de modelado bayesiano jerárquico como PYMC o el módulo de probabilidad de TensorFlow para ajustar distribuciones. Este enfoque aumentará la granularidad de las familias de distribuciones, permitiendo establecer parámetros de distribución por estación y por hora. Además, estas herramientas facilitan ajustes incluso con un volumen limitado de datos a este nivel de granularidad.

- 2.4 Adaptar el estudio a las nuevas características introducidas por BiciMad, como la actualización de la estructura de datos compartidos, los cambios en los modelos de bicicletas y estaciones, y la redefinición de la estructura tarifaria. Además, considerar el impacto de los periodos gratuitos en la demanda y las nuevas políticas municipales que afectan a los competidores y servicios alternativos.
- 2.5 Explorar nuevos algoritmos y métodos para modelar la popularidad de las estaciones, considerando variables socioeconómicas, orográficas, culturales (como lugares de interés) y urbanísticas (como la tipología de calles), entre otras.
- 2.6 Otra línea de investigación prometedora consiste en realizar un análisis estadístico robusto para aplicar la metodología propuesta a la predicción de fallos en la demanda insatisfecha. Esto requerirá tiempo para adaptar el modelo a las circunstancias actualizadas de BiciMad y, simultáneamente, recopilar un volumen sustancial de datos en tiempo real sobre el estado de las diversas estaciones.
- 2.7 Finalmente, sería deseable integrar el simulador con herramientas visuales, como dashboards dinámicos, para facilitar su adaptación e inclusión en la planificación operativa del servicio de BiciMad.

Como se puede concluir de las numerosas futuras líneas de investigación aquí mencionadas el ámbito de los BSS es un área de investigación prometedora y con potencial. Quedan varios estudios e investigaciones por realizar. Además hay que tener en cuenta la aplicabilidad práctica que todas ellas pueden tener al ser el transporte sostenible un asunto de interés global.

Apéndices

Apéndice A

Resumen Taxonómico de Algoritmos de Aprendizaje Automático (ML)

Tipo	Familia de Algoritmos	Algoritmo	Descripción
Probabilístico	Bayesian Methods	Bayesian Networks	Modelos probabilísticos gráficos que representan un conjunto de variables y sus dependencias condicionales mediante un grafo dirigido acíclico. Son utilizadas para realizar inferencia, modelar incertidumbre y tomar decisiones basadas en probabilidades. Pueden utilizarse tanto en aprendizaje supervisado como en no supervisado.
		Gaussian Processes	Modelos probabilísticos no paramétricos que describen distribuciones sobre funciones, permitiendo realizar predicciones con una medida de incertidumbre. Asumen que cualquier conjunto finito de puntos sigue una distribución gaussiana. Se suelen emplear para regresión y clasificación (supervisado) al ser capaces de modelar relaciones no lineales con incertidumbre.
		Hidden Markov Models (HMM)	Modelos probabilísticos que representan sistemas con estados ocultos que evolucionan en el tiempo según un proceso de Markov. Permiten modelar secuencias observadas y estimar las probabilidades de los estados ocultos detrás de ellas. Principalmente para aprendizaje no supervisado aunque también se puede usar en supervisado con secuencias etiquetadas.
	Latent Variable Models	Variational Autoencoders (VAEs)	Redes neuronales generativas que usan aproximación variacional para modelar variables latentes. Es decir, aprenden una representación latente de los datos mientras generan nuevos ejemplos similares a los originales. No supervisado.
		Restricted Boltzmann Machines (RBMs)	Modelos de energía que aprenden representaciones latentes. Su funcionamiento se basa en la definición de una función de energía que asigna un valor (similar a un coste o a una puntuación) a cada configuración posible de las variables visibles y ocultas del modelo. Se implementan por tanto con una red neuronal con dos capas: una visible y otra oculta. No supervisado principalmente.
Continúa en la siguiente página			

Tipo	Familia de Algoritmos	Algoritmo	Descripción
		Generative Adversarial Networks (GANs)	Redes neuronales generativas entrenadas mediante competencia entre dos redes, una que genera y otra que discrimina. la competencia se describe como que por un lado el generador mejora aprendiendo a engañar al discriminador, es decir, generando datos que el discriminador no pueda distinguir de los reales. Mientras que por otro el discriminador mejora aprendiendo a identificar con mayor precisión si los datos son reales o generados. No supervisado principalmente.
	Mixture Models and Curve Fitting	Gaussian Mixture Models (GMMs)	Modelos que representan la distribución de datos como una combinación de varias distribuciones gaussianas. Se usan para clustering (No supervisado).
		Dirichlet Processes	Herramientas estadísticas utilizadas en el modelado bayesiano no paramétrico. Son procesos estocásticos que permiten trabajar con una cantidad flexible e infinita de componentes en modelos de mezcla, sin necesidad de especificar previamente el número exacto de componentes (como sí ocurre con los GMMs) No supervisado.
	Probabilistic Inference Techniques	Maximum Likelihood Estimation (MLE)	Método de inferencia frecuentista que estima los parámetros del modelo maximizando la probabilidad de los datos observados dado el modelo. Supervisado y no supervisado.
		Maximum a Posteriori (MAP)	Método de inferencia probabilística bayesiana que tiene como objetivo estimar el parámetro θ de un modelo probabilístico que maximiza la distribución posterior $P(\theta \mid \text{datos})$. Supervisado y no supervisado.
		Markov Chain Monte Carlo (MCMC)	Técnicas para inferencia en modelos complejos mediante muestreo estocástico. No supervisado.
	Determinista	Linear Models	Linear Regression
Logistic Regression			Modelo que se emplea para clasificación (supervisado) y que utiliza la función logística/sigmoidea para calcular probabilidades de pertenencia de una observación a una clase específica. Pueden ser para clases binarias o multiclase.
Continúa en la siguiente página			

Tipo	Familia de Algoritmos	Algoritmo	Descripción
	Non-Linear Models	Support Vector Machines (SVMs)	es un modelo supervisado que se utiliza para clasificación y regresión. Su objetivo principal es encontrar la mejor línea (o hiperplano en espacios de mayor dimensión) que separe las diferentes clases en los datos.
		k-Nearest Neighbors (KNN)	Algoritmo basado en la proximidad de puntos en el espacio. Basado en asignar una clase (o un valor, en regresión) a un dato nuevo según las etiquetas de sus k-vecinos más cercanos Supervisado.
	Decision Tree-Based Models	Random Forests	Modelo supervisado basado en un conjunto de árboles de decisión entrenados en subconjuntos de datos, cuya combinación mejora la precisión y evita sobreajuste.
		Gradient Boosting Machines (GBMs)	Modelo supervisado que entrena secuencialmente árboles de decisión, cada uno corrigiendo los errores del anterior, para optimizar predicciones en regresión o clasificación.
	Neural Network-Based Models	Convolutional Neural Networks (CNNs)	Redes neuronales especializadas en datos estructurados en cuadrículas o mallas como por ejemplo son las imágenes. Supervisado.
		Recurrent Neural Networks (RNNs)	Redes neuronales para modelar datos secuenciales o temporales, manteniendo memoria de pasos previos mediante conexiones recurrentes. Supervisado.
	Clustering and Partitioning Methods	k-Means Clustering	Algoritmo no supervisado que agrupa datos en k clusters al minimizar la distancia de los puntos al centroide más cercano.
		Density-Based Spatial Clustering of Applications with Noise (DBSCAN)	Algoritmo de clustering basado en densidad que agrupa puntos cercanos entre sí y marca como ruido aquellos que no cumplen criterios de densidad mínima. No supervisado.

Tabla A.1: Resumen de los principales algoritmos de aprendizaje automático.

Apéndice B

Evolución de los mapas de estaciones de BiciMad

En este apéndice se presentan las imágenes que ilustran la evolución de los mapas de las estaciones de BiciMad. La Figura B.1 muestra la distribución de las estaciones originales al inicio del servicio en el año 2014.

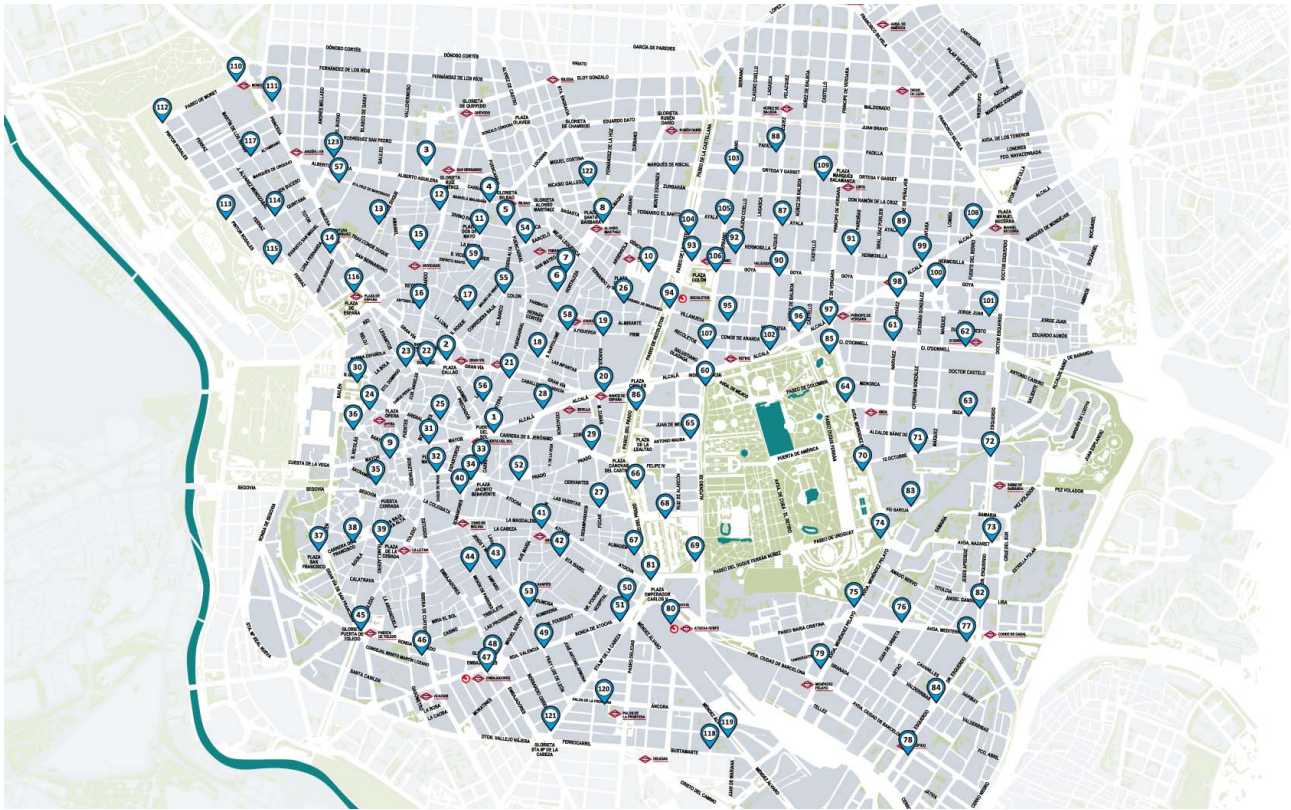


Figura B.1: Mapa con la localización e identificación de las estaciones de BiciMad en su inauguración en 2014. Fuente: web BiciMad

Seguidamente en la Figura B.2 se muestra el plano empleado en la licitación del contrato integral de movilidad a finales de 2014. En él se proyectan las nuevas estaciones que se desean incluir y se plantea reutilizar las existentes.

Entre los años 2016 y 2018 la página web de BiciMad presentaba el mapa que se muestra en la Figura B.3.

A continuación, las Figuras B.4 y B.5 muestran los planes de ampliación implementados en los años 2019 y 2020 respectivamente.

Finalmente, se incluyen la imagen de la propuesta de ampliación realizada en 2023 en la Figura B.6 y una sección de la representación del mapa actual de estaciones correspondiente a 2024 en la Figura B.7. Dado que éste último es extenso, cuenta con un formato dinámico y proporciona información adicional, como la ocupación en tiempo real, únicamente se presenta una sección del mismo. Se puede apreciar en la imagen la proliferación de estaciones que ha habido, teniendo que recurrir BiciMad a una representación dinámica donde se pueda hacer zoom.

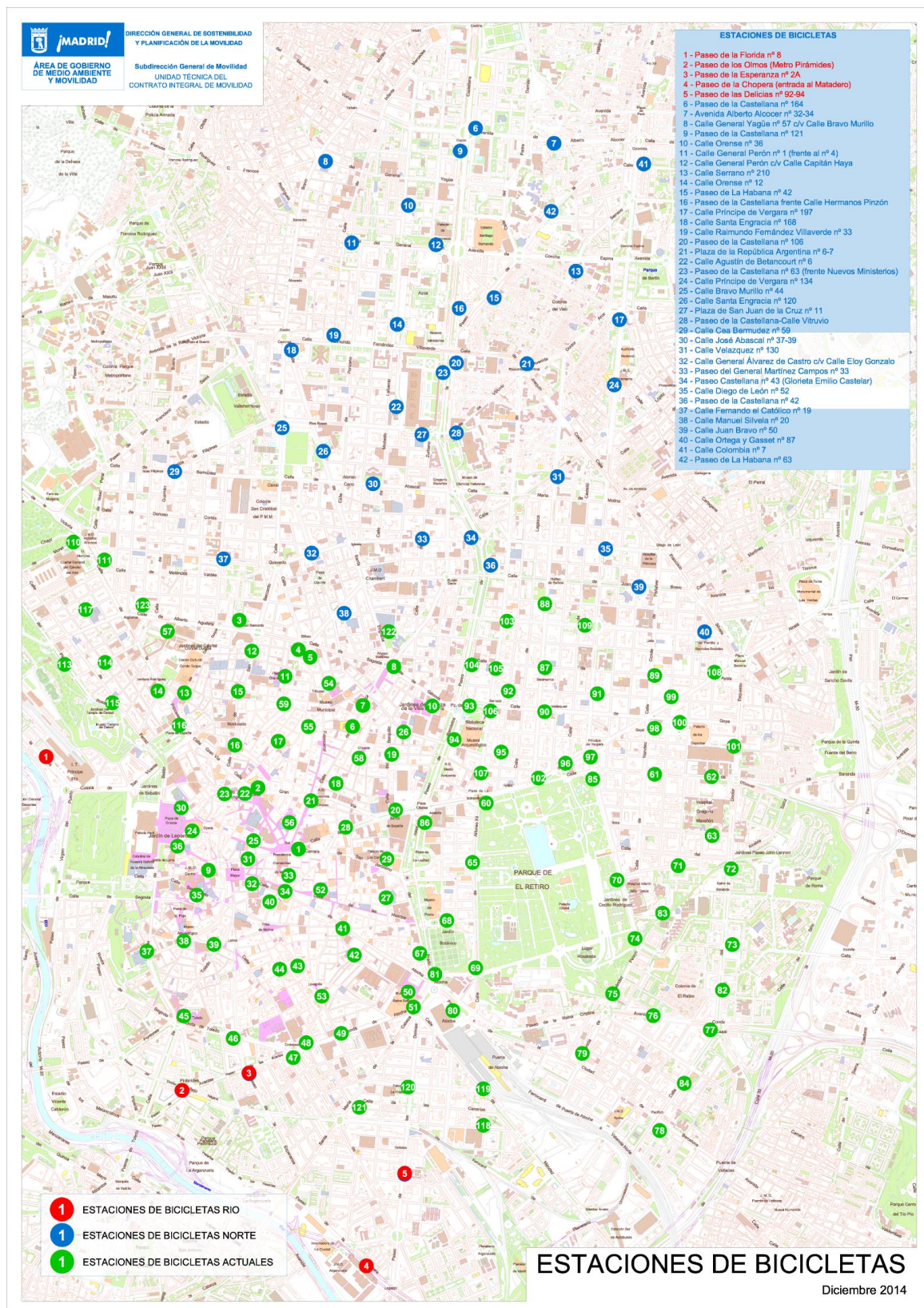


Figura B.2: Mapa de estaciones de BiciMad del Contrato Integral de Movilidad Diciembre 2014. Fuente: web BiciMad

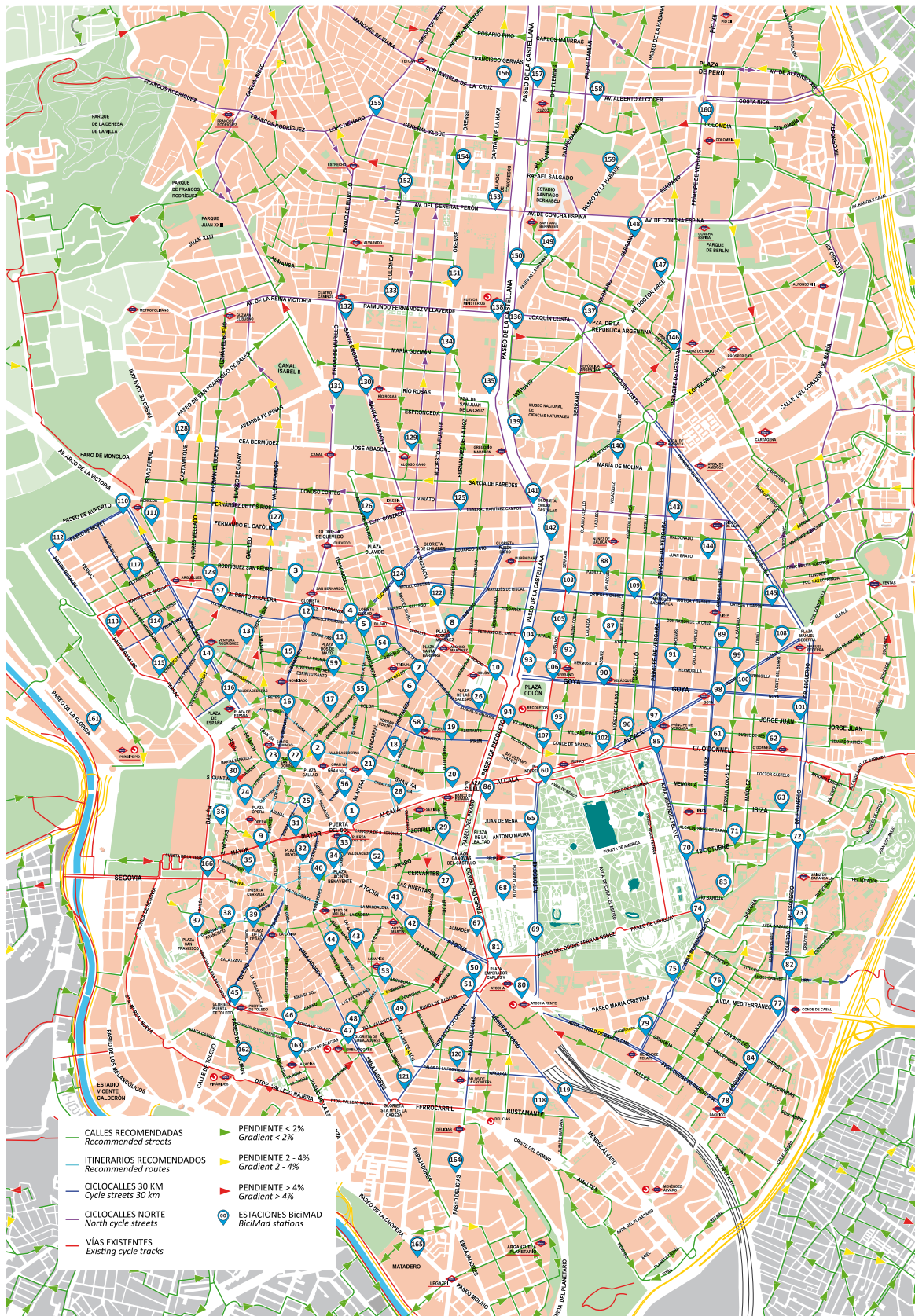


Figura B.3: Mapa estaciones de Bicimad 2015-2018. Fuente: web BiciMad

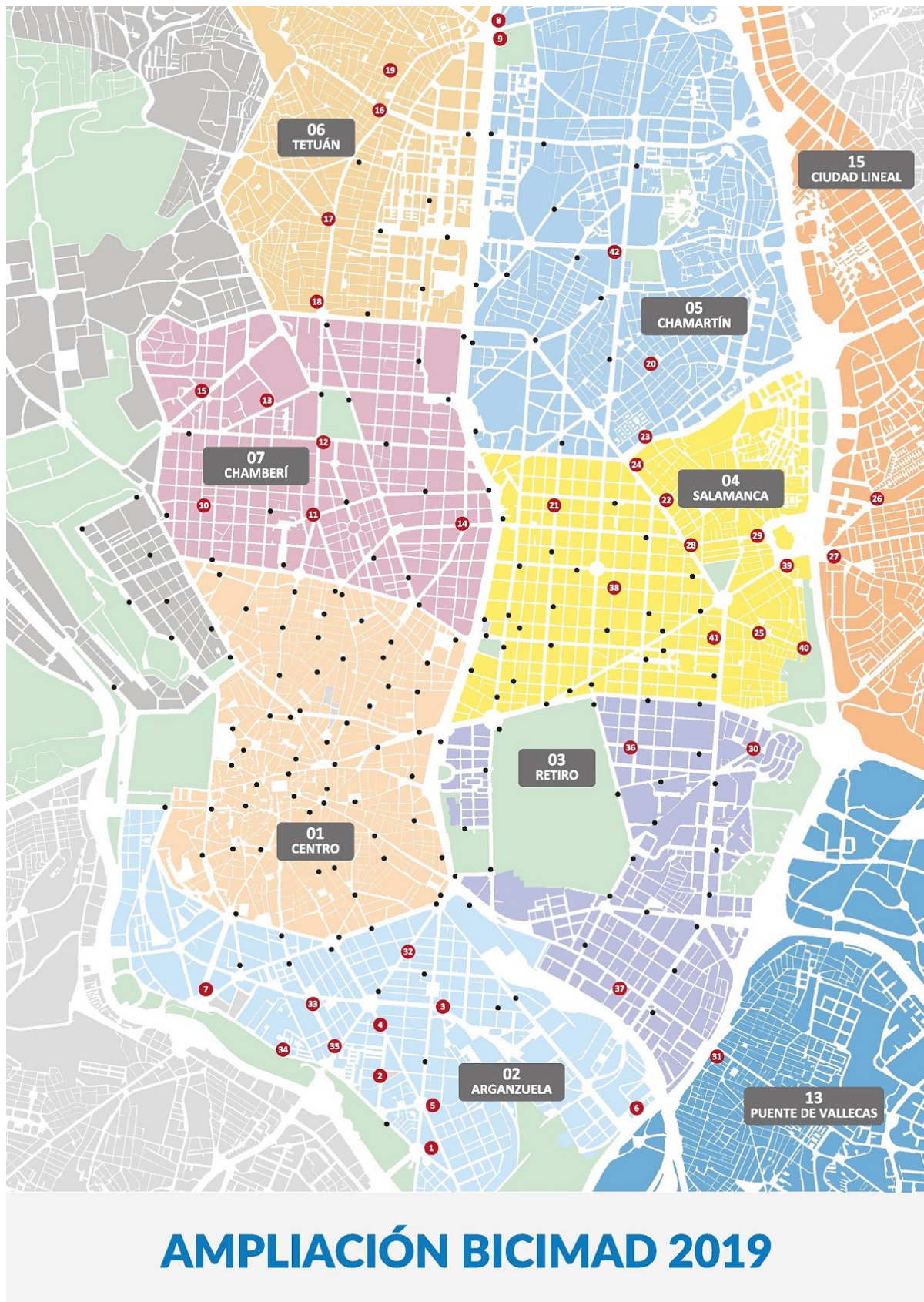


Figura B.4: Ampliación de estaciones de Bicimad en 2019. Fuente: web BiciMad

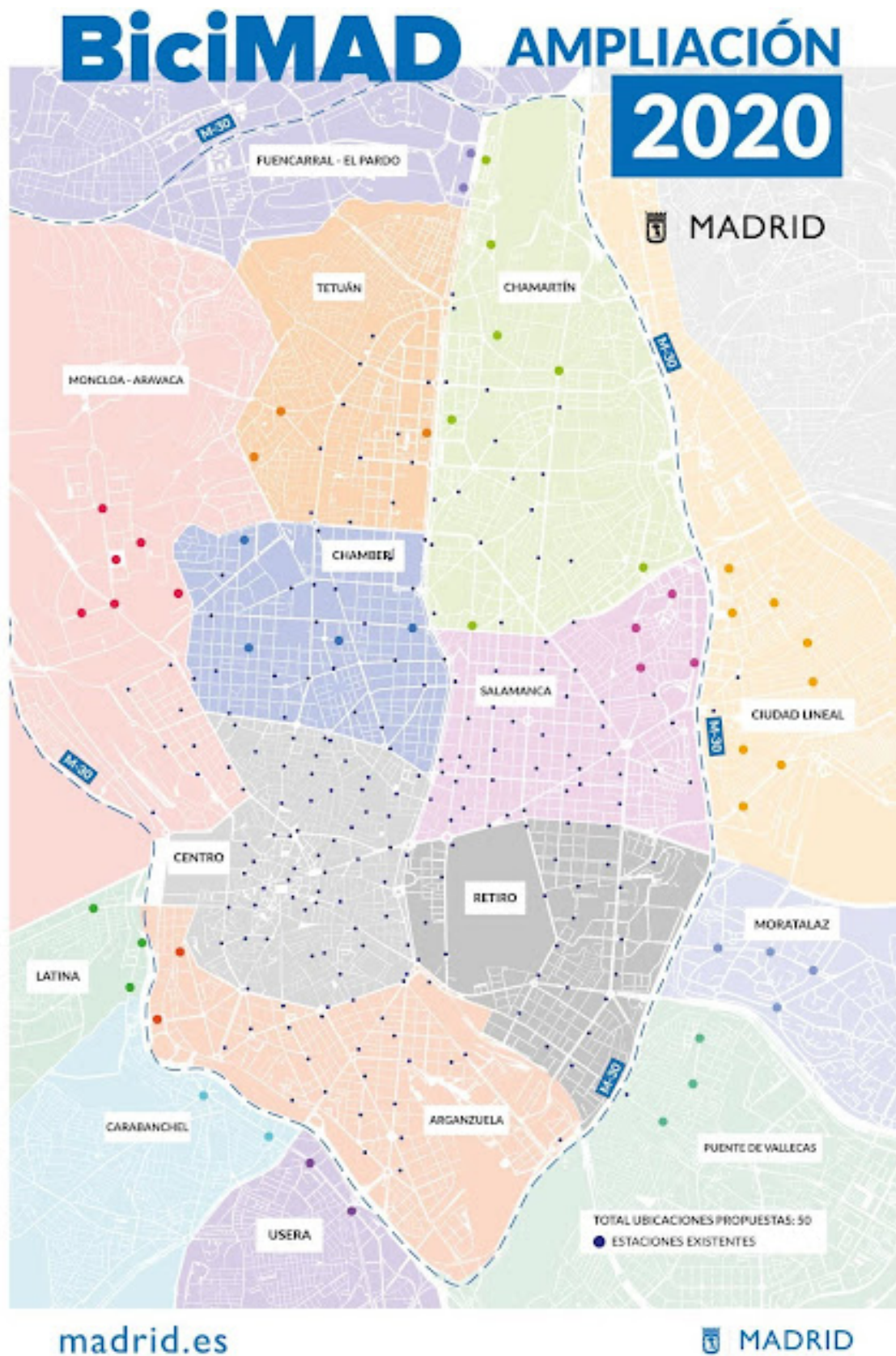


Figura B.5: Ampliación de estaciones de Bicimad en 2020. Fuente: web BiciMad

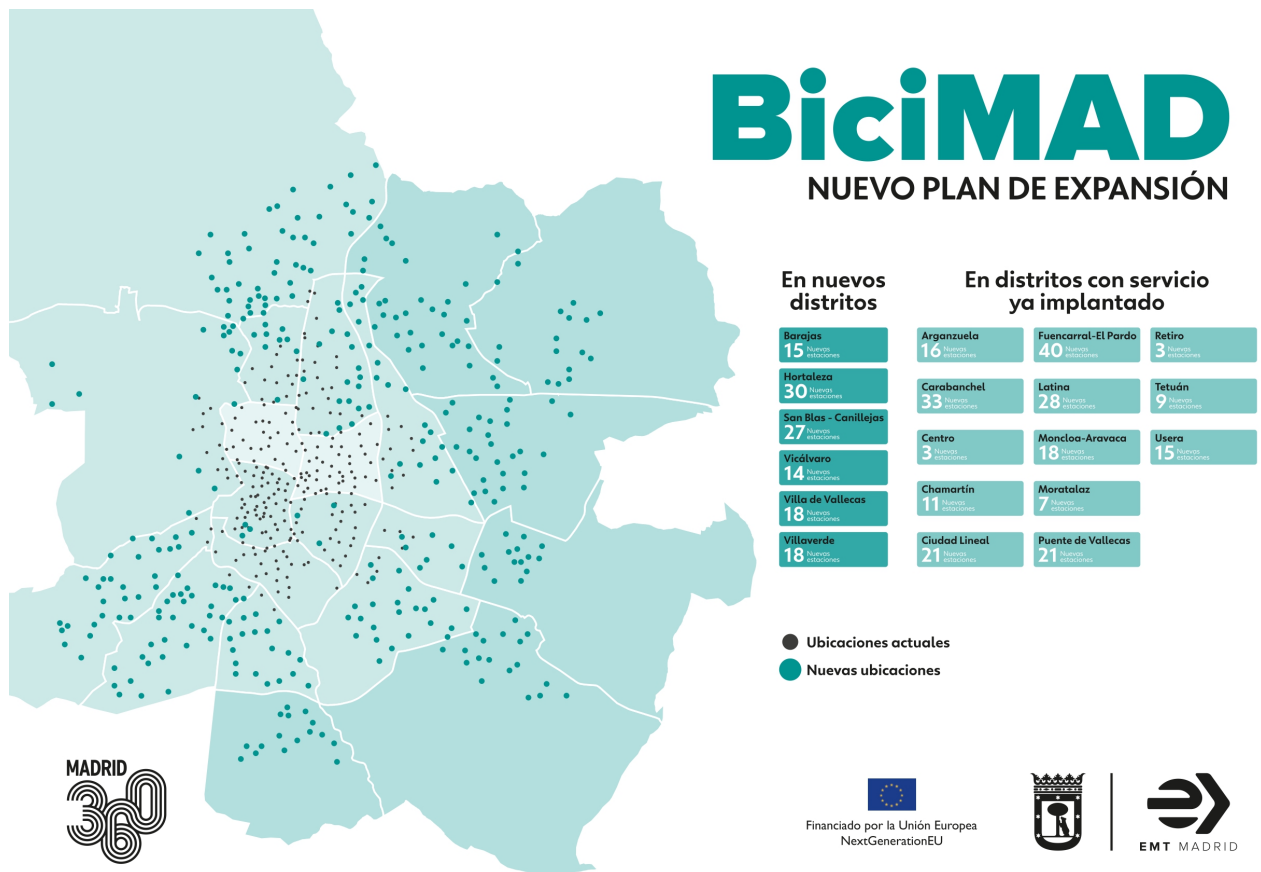


Figura B.6: Ampliación de estaciones de Bicimad en 2023. Fuente:web BiciMad

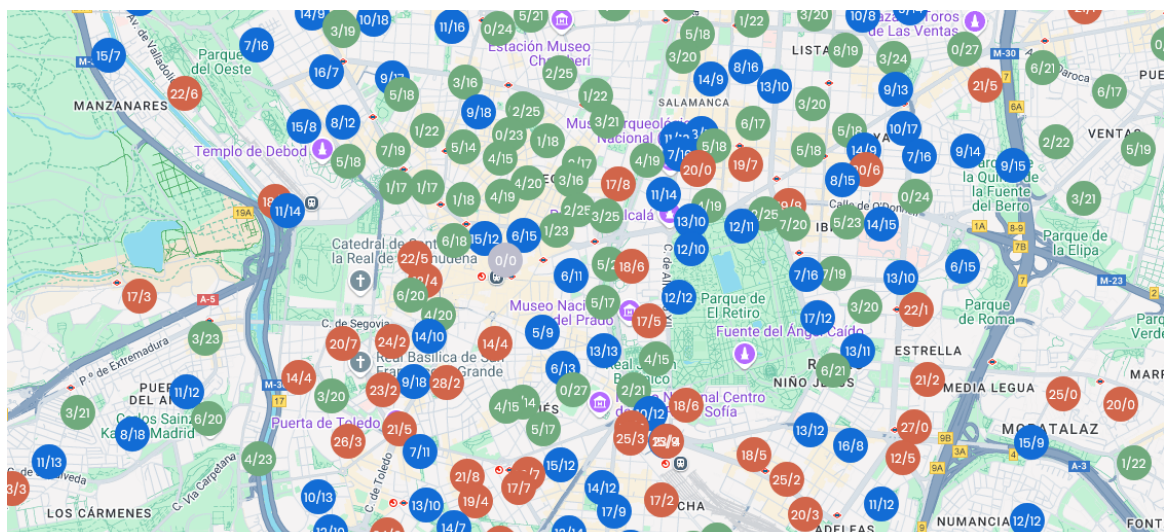


Figura B.7: Sección del mapa interactivo de estaciones de BiciMad 2024. Fuente: Elaboración propia captura de Mapa de BiciMad original

Apéndice C

Comparativa Evolutiva de BiciMad: Periodo 2018-2019 frente al Periodo 2023-2024

Concepto	BiciMad en 2018-2019	BiciMad en 2023-2024
Logo		
Nº de Estaciones	175	630
Fotografía de estación		
Nº de Bicicletas	2496	7735
Fotografía de bicicleta	 Se muestra inicialmente el modelo de BiciMad Clásico y debajo el del BiciMad Go.	

Concepto	Modelo: Booster Bikes BiciMad en 2018-2019	BiciMad en 2023-2024
Propiedades de las bicicletas	▪ Autonomía: 70 Km ▪ Peso: 22 Kg ▪ Material del cuadro: Acero inoxidable ▪ Velocidad Limitada a: 25 Km/h ▪ Pantalla: Monocromática para estado básico ▪ Sistema de anclaje: Mecánico en la estación ▪ Carga Máxima: 120 Kg ▪ Motor: En el buje trasero ▪ Sistema de transmisión: Mecánico de 3 marchas	▪ Modelo: E-Fit ▪ Autonomía: 70 Km pudiendo llegar a los 75 Km en ciertas condiciones. ▪ Peso: 19 Kg (más ligera) ▪ Material del cuadro: Aluminio ▪ Velocidad Limitada a: 25 Km/h ▪ Pantalla: Pantalla digital con mayor información básico ▪ Sistema de anclaje: Híbrido, es decir mecánico y magnético ▪ Carga Máxima: 130 Kg ▪ Motor: En el buje delantero ▪ Sistema de transmisión: Automático y continuo de 3 marchas ▪ Otras mejoras destacables: Sistema antipinchazo, frenos de tambor tipo rollerbrake, incorporan señal luminosa de frenado y mecanismo antirrobo integrado en el chasis, para mayor robustez y protección contra el vandalismo
	Nº de Distritos con servicio	821

Concepto	BiciMad en 2018-2019	BiciMad en 2023-2024
Sistema tarifario	Las tarifas de BiciMAD se en 2018 dividían en abonos anuales y tarjetas ocasionales. Los usuarios con abono anual pagan 25€ al año (o 15€ si tienen abono de transporte del Consorcio de Madrid) y 0,50€ por los primeros 30 minutos de uso, con 0,60€ por cada fracción adicional de 30 minutos. Las bonificaciones incluyen descuentos de 0,10€ por recoger bicicletas en estaciones excedentarias (ocupación > 70 %), devolverlas en estaciones deficitarias (ocupación < 30 %) o reservar un anclaje en destino (no siendo acumulables las distintas bonificaciones). Hay una penalización de 4€ por exceder el tiempo máximo de 2 horas, aplicada por hora o fracción. Para usuarios con tarjeta ocasional, la primera hora cuesta 2€ y la segunda 4€, con las mismas bonificaciones y penalizaciones que los abonados.	El sistema tarifario actual de BiciMad ofrece dos modalidades: contrato básico y tarifa plana. En el contrato básico, no hay cuota mensual y los primeros 30 minutos cuestan 0,50€, con un coste adicional de 3,00€ por cada fracción adicional de 30 minutos. Por otro lado, la tarifa plana tiene una cuota mensual de 10€, incluye los primeros 30 minutos gratis y un coste de 3,00€ por cada fracción adicional de 30 minutos.
Servicios sustitutivos y competencia directa	Existían concesiones tanto para sistemas de bicicletas compartidas sin base fija, como para patinetes eléctricos. Algunas de las empresas concesionarias era: Idbrik Spain, Bird, Boltest, Ride Dott y Lime.	El Ayuntamiento decide no renovar los permisos a las empresas de bicicletas compartidas sin base fija y de patinetes eléctricos compartidos. Por tanto BiciMad carece de competencia directa a finales de 2024.

Tabla C.1: Tabla comparativa resumen entre BiciMad de 2018 y BiciMad de 2024.

Apéndice D

Resumen de los artículos de revisión con sus principales características.

Año	Referencia	Título y enfoque principal	Métodos, conclusiones y principales aportaciones
2009	[DeMaio, 2009]	<i>Bike-sharing: History, Impacts, Models of Provision, and Future.</i> El artículo ofrece una visión histórica del bike-sharing desde los años 60 hasta la tercera generación. Analiza modelos de provisión, como los gestionados por gobiernos, agencias, universidades y empresas, destacando sus ventajas, desventajas y ejemplos. Finalmente, proyecta una cuarta generación con avances tecnológicos, como bicicletas eléctricas, estaciones solares y mejoras en distribución y modelos de negocio.	El artículo realiza una revisión histórica con análisis cualitativo de modelos operativos y costes, una comparativa de diferentes modelos de operación y una propuesta de mejoras futuras. Entre sus principales aportaciones cabe destacar: ■ Análisis de la evolución tecnológica. ■ Reconocimiento del impacto positivo en la sostenibilidad. ■ Valoración de beneficios para la salud. ■ Identificación de desafíos económicos a resolver como métodos de pagos, gestión de costes y control del vandalismo. ■ Sentar las bases para la aparición de la cuarta generación. En resumen, las conclusiones del artículo definen líneas de investigación clave para el estudio de los BSS.
2013	[Fishman et al., 2013]	<i>Bike share a synthesis of the literature.</i> El artículo sintetiza la literatura existente, hasta la fecha de su publicación, sobre bike-sharing. Adicionalmente realiza una comparación de implementaciones de BSS internacionales y analiza barreras y facilitadores, como la infraestructura y la regulación.	El estudio integra encuestas a usuarios, análisis de estudios previos y datos operativos para examinar los patrones de uso y la eficiencia en los BSS. También compara contextos normativos y culturales, como el efecto de la obligatoriedad del uso del casco, e identifica vacíos en la investigación, resaltando la falta de datos consistentes sobre distancias recorridas y la reducción del uso del automóvil. Las principales conclusiones del estudio son: ■ Los viajes usando BSS sustituyen a caminatas y transporte público. No tanto a viajes en coche. ■ BSS valorado principalmente por la accesibilidad y la utilidad en trayectos cortos. ■ BSS mejora la interconexión del transporte público, aportando una solución para el problema de ‘la última milla’.

Continúa en la siguiente página

Año	Referencia	Título y enfoque principal	Métodos, conclusiones y principales aportaciones
2015	[Fishman, 2015]	<i>Bikeshare: A Review of Recent Literature.</i> Este artículo de revisión supone una actualización del anterior del mismo autor centrado principalmente en analizar las tendencias globales en el uso de BSS y su impacto en la movilidad urbana. Para ello examina las barreras y motivaciones de los usuarios para el uso de este tipo de servicios. Finalmente pone el foco en la investigación futura relacionándola con áreas como la seguridad, la sostenibilidad y la integración tecnológica.	El artículo realiza una revisión sistemática de literatura sobre BSS desde 2013, empleando bases de datos académicas como Scopus y Google Scholar con palabras clave terminológicamente fragmentadas. Las principales conclusiones del estudio son: ■ Los BSS promueven la movilidad sostenible, pero su impacto en la reducción de emisiones y reducción del uso de otros vehículos es limitado. ■ Factores como la proximidad de las estaciones y el coste del servicio son determinantes clave para una adopción exitosa del mismo. ■ Reconocimiento del rol de la tecnología en los sistemas de tercera y cuarta generación como clave para el crecimiento global. Principalmente la incorporación de bicicletas eléctricas o sistemas sin necesidad de anclaje.
2015	[Ricci, 2015]	<i>Bike Sharing: A Review of Evidence on Impacts and Processes of Implementation and Operation.</i> Esta revisión primeramente evalúa los impactos sociales, ambientales y económicos de los BSS. Seguidamente analiza los procesos de implementación y operación de este tipo de servicios con el fin de identificar buenas prácticas y las barreras comunes. Finalmente compara las características demográficas de los usuarios y los patrones de uso.	Se realiza un análisis crítico de la literatura, con foco principalmente en los estudios que poseen una metodología robusta y disponen de datos empíricos, centrándose en regiones clave de América, Europa y Asia. Las principales conclusiones del estudio son: ■ La recolección y análisis de datos resulta esencial para realizar mejores evaluaciones del servicio. ■ Reconocimiento del costo y complejidad del proceso de reequilibrio, y por extensión la necesidad de innovación en los métodos empleados para implementar ese reequilibrio, ■ La relación entre la ubicación y la demanda. Resaltando que dependiendo de la ubicación de las estaciones así serán las necesidades de reequilibrio de bicicletas.
Continúa en la siguiente página			

Año	Referencia	Título y enfoque principal	Métodos, conclusiones y principales aportaciones
2019	[Si et al., 2019]	<i>Mapping the bike sharing research published from 2010 to 2018: A scientometric review.</i> El estudio utiliza un enfoque cientométrico para analizar la investigación sobre BSS entre los años 2010 y 2018.	Emplea herramientas como CiteSpace, ayudando a crear mapas visuales de conocimiento mediante el análisis de co-citación de documentos, palabras clave y agrupamiento por temas. Las principales aportaciones de la revisión son: ■ Se identificaron cinco áreas de investigación principales: factores y barreras, optimización de sistemas, comportamiento e impacto, seguridad y salud, y economía compartida. ■ Introduce las implementaciones de BSS sin estación base y que emplean nuevas tecnologías (quinta generación). ■ Enumera las principales instituciones que contribuyen en la investigación de los BSS.
2021	[Vallez et al., 2021]	<i>Challenges and Opportunities in Dock-Based Bike-Sharing Rebalancing: A Systematic Review</i> Se trata del primer paper de nuestra autoría que forma parte de esta tesis. El enfoque de esta revisión fue triple: Primeramente temporal incorporando las nuevas investigaciones entre 2018 y 2020. Segundo terminológico, reduciendo la fragmentación de la nomenclatura específica de BSS en la literatura, Finalmente centrado en el concepto del rebalanceo.	Búsqueda de papers empelando para ello un Tesauro de términos relacionados con BSS para evitar la dispersión de investigaciones debido a la fragmentación terminológica. Seguidamente un filtrado temporal entre los aós 2010 y 2020, uno terminológico para centrar el foco en el problema del rebalanceo y por último uno de subdivisión en los BSS para obtener solo los papers relacionados con BSS que tengan estaciones (<i>dock-based</i>) Las principales aportaciones de nuestro primer paper son: ■ Tesauro que ayuda a unificar la fragmentación terminológica. ■ Explicación del problema del rebalanceo como la suma de un subproblema de predicción de la demanda del servicio y otro de optimización de rutas de distribución de bicicletas. ■ Recopilación y clasificación de literataura relacionada con el problema del rebalanceo en BSS. ■ Identificación de nuevas líneas de investigación y preguntas abiertas en el problema de rebalanceo de los BSS.
Continúa en la siguiente página			

Año	Referencia	Título y enfoque principal	Métodos, conclusiones y principales aportaciones
2022	[Zhou et al., 2022]	<i>Review of bike-sharing system studies using bibliometrics method</i> Enfoque cientimétrico combinado con un análisis bibliométrico.	El método consiste en recolectar datos de la base Web of Science (WoS) con un filtrado de documentos relevantes, seguido de análisis y visualización mediante las herramientas VOSviewer y CiteSpace para estudiar redes de co-citación y tendencias de investigación. Finalmente, se clasifican los temas en áreas. Las principales aportaciones de esta revisión son: ■ Muestra un importante crecimiento de investigaciones en el área de los BSS. ■ Identifica como áreas de investigación clave el problema del rebalanceo de bicicletas, las características espaciotemporales y la predicción de demanda. ■ Por el enfoque empleado han descubierto redes colaborativas y áreas emergentes en la investigación, siendo capaces también de concretar las instituciones y zonas geográficas que más aporte realizan al campo.
2023	[Zhang et al., 2023]	<i>Relocation-related problems in vehicle sharing systems: A literature review.</i> El enfoque seguido consiste en un análisis de los problemas de reubicación de vehículos en sistemas compartidos (car-sharing, bike-sharing y scooter-sharing.	El método incluyó una revisión sistematizada de la literatura, la clasificación de estrategias en niveles jerárquicos, el análisis de modelos matemáticos empleados, la comparación de soluciones en diversos sistemas de car-sharing, bike-sharing y scooter-sharing, y finalmente la identificación de gaps en la investigación. Hay que destacar las siguientes conclusiones: ■ La reubicación dinámica de vehículos es crítica para resolver desequilibrios en los sistemas. El uso de datos en tiempo real puede mejorar la precisión de las estrategias de reubicación. ■ Los métodos de reubicación basados en incentivos al usuario son menos costosos pero menos efectivos que los basados en acciones por parte del operador del servicio.

Continúa en la siguiente página

Año	Referencia	Título y enfoque principal	Métodos, conclusiones y principales aportaciones
2023	Teusch et al. [2023]	<i>A Systematic Literature Review on Machine Learning in Shared Mobility</i> El enfoque está centrado en el análisis de las técnicas de <i>Machine Learning</i> en la gestión de sistemas de vehículos compartidos. Por tanto abarca tanto a las técnicas aplicadas la parte del modelado de la demanda como a la parte de la optimizaciónn de la reubicación de los vehículos. Ambos son subproblemas del ampliamente reconocido problema del rebalanceo de vehículos en estos sistemas.	El método empleado parte de una recopilación de papers entre los años 2012 y 2022. A continuación realiza un náalisis de métodos de <i>ML</i> aplicados en <i>car-sharing</i> , <i>bike-sharing</i> y otros modos de movilidad compartida. Tras identificar los conjuntos de datos utilizados para validar los diferentes métodos, se construye un marco conceptual de técnicas <i>ML</i> aplicadas a la movilidad compartida. Finaliza señalando las áreas de investigación no cubiertas aun por la literatura. Entre los principales resultados que aporta, caben destacar: ▪ <i>ML</i> es un método efectivo para la predicción de demanda, optimización de rutas y estrategias de reubicación. ▪ La mayoría de estudios se centran en aplicaciones operativas más que en decisiones estratégicas a largo plazo. ▪ La disponibilidad limitada de conjuntos de datos abiertos sigue siendo un desafío significativo.
2021	Albuquerque et al. [2021]	<i>Machine Learning Approaches to Bike-Sharing Systems: A Systematic Literature Review</i> En esta revisión el enfoque es aún más específico que la anterior, ya que aunque se centra igualmente en el análisis de la aplicación de técnicas de <i>Machine Learning</i> , su ámbito se reduce a los <i>BSS</i> .	El método empleado parte de una recopilación de papers entre los años 2015 y 2019. Seguidamente realiza una división por tipo de técnica empleada entre clasificación y predicción. Se identifican influencias de factores externos como condiciones climáticas, infraestructura y comportamiento del usuarios. Y finalmente se identifican gaps en las diversas implementaciones de <i>Machine Learning</i> . Destacan como principales conclusiones de este artículo las siguientes: ▪ Los algoritmos de predicción como Random Forest y Redes Neuronales son los más utilizados para la demanda de <i>BSS</i>. ▪ Las técnicas de clasificación ayudan a segmentar patrones de uso según factores temporales y demográficos. ▪ Los factores externos, como el clima, afectan significativamente la precisión de los modelos. ▪ Existe un gap para evolución de la investigación que consiste en la implementación de soluciones híbridas, combinando múltiples algoritmos de <i>Machine Learning</i>.

Tabla D.1: Resumen de los principales artículos de revisión sobre sistemas de bicicletas compartidas

Apéndice E

**Resumen de artículos sobre el rebalanceo en los
BSS y sus principales características.**

Referencia	Región / Ámbito	Estrategia / Algoritmo	Contribución Clave
Nair & Miller-Hooks, 2011	America	Stochastic MIP	Very simplistic model for BBS but valid for
Benchimol et al., 2011	Europe	Approximation algorithms and polynomial algorithm	Treat problem as alternative of the classic traveling salesman problem (TSP)
Wood, Slingsby, & Dykes, 2011	Europe	Visualization Techniques	Identifies that service demand and implementation of BSS affect to rebalancing strategy. Any solution will require a kind of study of both characteristics.
Chemla, Meunier, & Wolfler Calvo, 2012	Europe	A branch-and-cut algorithm applied to an integer program with several constraints.	First time metaheuristic approach is applied to this problem.
Labadi, Benarbia, Hamaci, & Darcherif, 2012	Europe	Petri nets	First study to include Petri nets to simulate and validate models of this transportation service.
Lin & Chou, 2012	Asia	GIS information + farthest insertion and nearest neighbor algorithms	Use GIS information.
Contardo, Morency, & Rousseau, 2012	America	Column generation and Benders decomposition applied to a model based on arc-flow	Develop a scalable methodology that provides lower and upper bounds in short computing times.
Raviv et al., 2013	Asia	Mixed integer linear program (MILP)	Use of MILP models.
Rainer-Harbach, Papazek, Hu, & Raidl, 2013	Europe	Variable Neighborhood Search (VNS) with an embedded Variable Neighborhood Descent (VND) plus use of greedy heuristic, a maximum flow calculation, and linear programming	Experimental results that show that VNS performs well and scales much better than other MIP approaches.
Raidl, Hu, Rainer-Harbach, & Papazek, 2013	Europe	Variable Neighborhood Search (VNS) combined with Greedy Heuristics (GH), maximum flow approach for the monotonic case (MF-MC) and maximum flow-based, general case (MF-GC)	Experimental results indicate that this combined approach yields significantly better results than the original variable neighborhood search.
<i>Continúa en la siguiente página</i>			

Referencia	Región / Ámbito	Estrategia / Algoritmo	Contribución Clave
Nair, Miller-Hooks, Hampshire, & Bušić, 2013	America	Stochastic MIP	They show the efficacy of generating redistribution strategies using stochastic information applied to a large-scale BSS implementation.
J. R. Lin, Yang, & Chang, 2013	Asia	Heuristics	A hub location inventory model.
Di Gaspero, L., Rendl, A., & Urli, T. 2013	Europe	Constraints programming speed-up with Large Neighborhood Search (LNS)	Not outperforms current solutions but allows a more general formulation of a BSS.
Erdoğan, G., Laporte, G., & Calvo, R. W. 2014	Europe	A branch-and-cut algorithm applied to an integer program	Consider lower and upper bounds.
Kloimüller, C., Papazek, P., Hu, B., & Raidl, G. R. 2014	Europe	Variable Neighborhood Search (VNS) combined with Greedy Heuristics (GH)	Improvement of the paper Raidl, Hu, Rainer-Harbach, & Papazek, 2013 done for dynamic scenario.
Regue, R., & Recker, W. 2014	America	Framework based on four models	Proactive approach (repositioning is done before unbalances happen) opposite to the traditional reactive approach.
Pfrommer, J., Warrington, J., Schildbach, G., & Morari, M. 2014	Europe	Solution based on a mix of intelligent routing plan plus dynamic rewards incentives to the users	First paper that evaluates the approach of using customer incentives (User-based approach).
Dell'Amico, M., Hadjicostantinou, E., Iori, M., & Novellani, S. 2014	Europe	Four proposed ways, all compared and solved with a branch-and-cut algorithm	Evaluation based on computational benchmarks.
Beecham, Wood, & Bowerman, 2014	Europe	Visual classification techniques	Identification of imbalances sources (mornings vs evenings, travels originated inside London vs outside ones).
Ho, S. C., & Szeto, W. Y. 2014	Europe/Asia	Iterated tabu search heuristic	High-quality solutions with short computation times.
Forma, I. A., Raviv, T., & Tzur, M. 2015	Europe	3-step heuristic method	Reduce the routing problem by clustering techniques.
Erdoğan, G., Battarra, M., & Calvo, R. W. 2015	Europe	Algorithm based on a linear programming solution	Tested on benchmark instances of the literature.
Alvarez-Valdes, R., Belenguer, J. M., Benavent, E., Bermudez, J. D., Muñoz, F., Vercher, E., & Verdejo, F. 2016	Europe	Mathematical-solution based to obtain the predictions and heuristics to decide the routing part.	Identify the two parts that all solutions to the repositioning problem have. Provide a solution that merges both and suggests extending research to dynamic solutions.
<i>Continúa en la siguiente página</i>			

Referencia	Región / Ámbito	Estrategia / Algoritmo	Contribución Clave
Dell, M., Iori, M., Novellani, S., & Stützle, T. 2016	Europe	New heuristic with local searches and adapting it with a branch-and-cut technique	An improvement in computational time.
Li, Y., Szeto, W. Y., Long, J., & Shui, C. S. 2016	Asia	Mixed-linear programming problem and solving it with a hybrid genetic algorithm	They consider Scenario of having different types of bikes.
Cruz, F., Subramanian, A., Bruck, B. P., & Iori, M. 2016	America, Europe	Iterated Local Search (ILS) based heuristic	Comparison with the solutions proposed by (Erdogan et al., 2015) and (Chemla et al., 2012).
Liu, J., Sun, L., Chen, W., & Xiong, H. (2016	America, Asia	1 a Meteorology Similarity Weighted K-Nearest-Neighbor (MSWK) regressor to predict the demand based on historical data, 2 a mixed integer nonlinear programming (MINLP) trying to minimize travel distance of routing vehicle, and 3 a AdaptiveCapacity Constrained K-centers Clustering (AdaCCKC) algorithm to create clusters of stations	Large scale rebalance optimization
Zhang, D., Yu, C., Desai, J., Lau, H. Y. K., & Srivathsan, S. 2016	Asia	Mixed-integer problem solved with a novel heuristic algorithm.	Significant reduction in rejected user requests compared with existing methodology
Szeto, W. Y., Liu, Y., & Ho, S. C. 2016	Europe, Asia	Chemical reaction optimization (CRO)	Improve the slution quality reducing computing time
Ho, S. C., & Szeto, W. Y. 2017	Europe, Asia	Heuristic based on large neighborhood search, with tabu search and various insertion and removal operators	Improve performance
Shui, C. S., & Szeto, W. Y. (2017)	Asia	Artificial bee colony (EABC) algorithm and a route truncation heuristic	Focus objective in CO2 emissions reduction
<i>Continúa en la siguiente página</i>			

Referencia	Región / Ámbito	Estrategia / Algoritmo	Contribución Clave
Chiariotti, F., Pielli, C., Zanella, A., & Zorzi, M. 2018	Europe	Combination of Birth-Death Processes and graph theory	Clear benefits of dynamic vs static approach.
Elhenawy, M., & Rakha, H. 2017	America	Game theory : based on the deferred acceptance algorithm	An approach that some times is better than known solution and other times close to it but with a better computational performance.
Schuijbroek, J., Hampshire, R. C., & Van Hoes, W. J. 2017	Europe America	Cluster-first route-second heuristic ,	First paper that unifies station inventory level constraints with vehicle routing for static rebalancing
Bulhões, T., Subramanian, A., Erdoğan, G., & Laporte, G. 2018	America/Europe	Integer programming problem solved with a branch-and-cut algorithm	Relationship between number of revisits, number of vehicles and capacity of them
Szeto, W. Y., & Shui, C. S. 2018	Asia	Enhanced artificial bee colony (EABC) algorithm	Better performance and higher quality solutions than CPLEX. A set of best practices that have been demonstrated.
Legros, B. 2019	Europe	Markov decision process	Provide a solution with simpler prioritization rules and better performance.
Dell'Amico, M., Iori, M., Novellani, S., & Subramanian, A. 2018	Europe	Heuristic with Deterministic Equivalent Program, L-Shaped methods, and branch-and-cut algorithms	Treat stochastic version of the problem.
Li, Y., Zheng, Y., & Yang, Q. 2018, July	Asia	Spacio-temporal reinforcement learning based on clustering algorithms	Dynamic approach based on clustering and tested with real dataset, providing a better solution than the used one.
You, P. S. 2019	Asia	Constrained nonlinear mixed integer programming model.	Better solution than GAMS/CEPLEX and a mixed approach (genetic algorithm plus linear programming)
Tang, Q., Fu, Z., & Qiu, M. 2019	Asia	Bi-level model solve with iterated local search and tabu search algorithms	Better performance and demonstrates some insights that affects the cost of repositioning
Wang, S., & Wu, R. 2019	Asia	Heuristics based mathematical algorithms	Include the concept of unusable bikes and stations into the variables of the problem.
Warrington, J., & Ruchti, D. 2019	America Europe	A two-stage stochastic approximation	Demonstrate a better performance than real options and provide a solution for other vehicle-sharing systems”;
<i>Continúa en la siguiente página</i>			

Referencia	Región / Ámbito	Estrategia / Algoritmo	Contribución Clave
Brinkmann, J., Ulmer, M. W., & Mattfeld, D. C. 2020	Europe	Stochastic-dynamic inventory routing problem modeled as a Markov decision process	Outperforms benchmark policies from the literature
Tang, Q., Fu, Z., Zhang, D., Guo, H., & Li, M. 2020	Asia	A two-stage stochastic programming model solved with simulated annealing algorithm	Focus objectives in costs reduction
Vishkaei, B. M., Mahdavi, I., Mahdavi-Amiri, N., & Khorram, E. 2020	Asia	A Jackson network solved with a genetic algorithm	Acceptable improvement and provide advice regarding fleet size and the BSS capacity
Zhang, D., Xu, W., Ji, B., Li, S., & Liu, Y. 2020	Asia	Linear programming model solved with adaptive tabu search algorithm	Better performance compared with tabu search (TS) and variable neighborhood search (VNS).
Jia, Y., Xu, Y., Yang, D., & Li, J. 2020	Asia	A bi-objective integer-mixed programming model, using a multi-start multi-objective particle swarm optimization (MS-MOPSO) algorithm to solve it.	Introduce the concept of balance interval.
Lu, Y., Benlic, U., & Wu, Q. (2020). hline	Asia	Memetic algorithm	Two to six times faster than existing heuristics

Tabla E.1: Resumen de artículos relacionados con los diversos enfoques empleados en la literatura para abordar el problema del rebalanceo de bicicletas en los BSS

Apéndice F

Análisis de viajes diarios según temperatura, tipo de día y precipitaciones

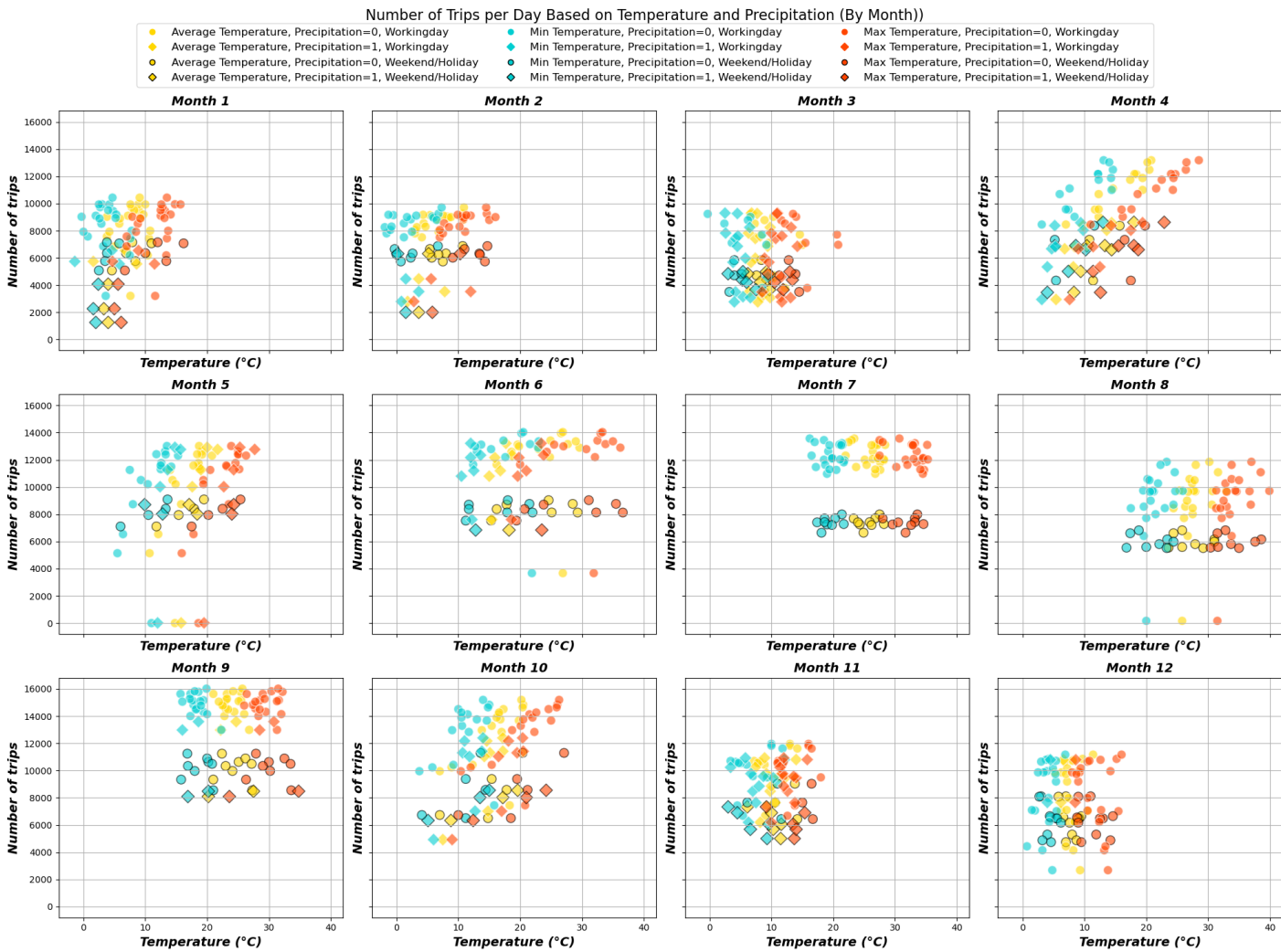


Figura F.1: Visualización del análisis del número de viajes diarios en relación con las temperaturas máxima, mínima y promedio, considerando además si el día correspondía a un festivo/fin de semana o un día laboral, así como la presencia o ausencia de precipitaciones. 1

```

OLS Regression Results
=====
Dep. Variable:          NUM_TRIPS      R-squared:                0.455
Model:                  OLS            Adj. R-squared:           0.450
Method:                 Least Squares   F-statistic:              97.80
Date:                   Wed, 20 Nov 2024 Prob (F-statistic):        4.65e-46
Time:                   11:08:23        Log-Likelihood:           -3268.2
No. Observations:       356            AIC:                     6544.
Df Residuals:           352            BIC:                     6560.
Df Model:                3
Covariance Type:        nonrobust
=====
               coef      std err          t      P>|t|      [0.025      0.975]
-----
Intercept      7604.3091    317.910     23.920     0.000     6979.068     8229.551
weekend_flag  -3063.9725    278.140    -11.016     0.000    -3610.997    -2516.948
TEMP_AVE       173.8699     16.374     10.619     0.000      141.668      206.072
PREC_CLASS    -1157.5658    303.856     -3.810     0.000    -1755.168    -559.964
=====
Omnibus:            86.676   Durbin-Watson:           1.990
Prob(Omnibus):      0.000   Jarque-Bera (JB):        209.850
Skew:               -1.187   Prob(JB):                 2.70e-46
Kurtosis:           5.918   Cond. No.                  52.5
=====

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```

Figura F.2: Regresión Lineal de Viajes Diarios en función de las Variables Meteorológicas y del Tipo de Día

Apéndice G

Parámetros de la distribución Binomial
Negativa para viajes por hora según tipo de día
y lluvia

w	r	hora	p	n	media	varianza	Desv. Est.
0	0	0	0.021	6.049	275.946	12864.548	113.422
0	0	1	0.032	5.332	161.049	5025.522	70.891
0	0	2	0.045	4.456	95.142	2126.614	46.115
0	0	3	0.059	3.662	57.891	973.048	31.194
0	0	4	0.082	3.618	40.295	489.028	22.114
0	0	5	0.140	7.444	45.831	327.999	18.111
0	0	6	0.174	17.881	85.114	490.253	22.142
0	0	7	0.026	9.906	368.527	14078.988	118.655
0	0	8	0.015	13.416	861.043	56124.770	236.907
0	0	9	0.023	17.398	754.511	33476.065	182.965
0	0	10	0.056	22.698	385.962	6949.058	83.361
0	0	11	0.078	29.639	349.355	4467.180	66.837
0	0	12	0.074	30.297	376.541	5056.358	71.108
0	0	13	0.057	29.277	488.093	8625.243	92.872
0	0	14	0.031	24.467	753.749	23974.199	154.836
0	0	15	0.021	16.499	783.239	37966.215	194.849
0	0	16	0.036	23.753	630.858	17385.980	131.856
0	0	17	0.033	22.605	659.377	19893.131	141.043
0	0	18	0.014	12.258	864.082	61773.562	248.543
0	0	19	0.014	12.663	870.705	60739.541	246.454
0	0	20	0.019	14.158	750.353	40518.555	201.292
0	0	21	0.027	16.421	586.191	21511.991	146.670
0	0	22	0.028	11.641	398.978	14073.728	118.633
0	0	23	0.027	9.377	336.426	12406.878	111.386
0	1	0	0.021	4.639	218.852	10543.830	102.683
0	1	1	0.037	4.853	127.574	3481.356	59.003
0	1	2	0.048	3.811	75.093	1554.751	39.430
0	1	3	0.069	3.467	46.833	679.546	26.068
0	1	4	0.103	3.822	33.278	323.015	17.973
0	1	5	0.186	8.996	39.296	210.949	14.524
0	1	6	0.142	12.233	74.000	521.630	22.839
0	1	7	0.025	7.407	291.019	11725.018	108.282
0	1	8	0.010	7.701	731.463	70210.360	264.972
0	1	9	0.010	6.740	636.204	60691.347	246.356
0	1	10	0.018	5.318	289.481	16045.916	126.672
0	1	11	0.018	4.558	249.815	13940.188	118.069
0	1	12	0.016	4.402	277.148	17727.422	133.144
0	1	13	0.012	4.682	370.259	29652.377	172.199
0	1	14	0.009	4.996	578.352	67531.043	259.867
0	1	15	0.008	4.546	559.685	69466.216	263.564
0	1	16	0.011	5.276	460.315	40622.808	201.551
0	1	17	0.011	5.375	504.130	47791.483	218.613
0	1	18	0.006	3.965	657.722	109755.793	331.294
0	1	19	0.006	3.884	622.852	100493.311	317.007
0	1	20	0.008	4.494	526.222	62145.321	249.290
0	1	21	0.013	5.221	405.648	31924.117	178.673
0	1	22	0.018	5.204	286.111	16015.654	126.553
0	1	23	0.025	5.778	229.130	9315.335	96.516
1	0	0	0.031	11.143	349.538	11313.915	106.367
1	0	1	0.031	10.365	326.707	10624.524	103.075
1	0	2	0.034	10.116	290.580	8637.552	92.938
1	0	3	0.045	10.078	211.598	4654.167	68.221
1	0	4	0.069	11.135	149.524	2157.396	46.448
1	0	5	0.109	14.173	116.402	1072.411	32.748
1	0	6	0.149	17.064	97.366	652.939	25.553
1	0	7	0.179	13.867	63.817	357.515	18.908
1	0	8	0.092	9.619	95.085	1035.029	32.172
1	0	9	0.039	6.428	156.378	3960.479	62.932
1	0	10	0.042	9.925	224.268	5292.026	72.746
1	0	11	0.046	14.578	299.256	6442.532	80.265
1	0	12	0.040	15.047	356.805	8817.767	93.903
1	0	13	0.033	14.723	432.049	13110.168	114.500
1	0	14	0.034	15.588	441.585	12951.243	113.804
1	0	15	0.047	16.698	336.976	7137.170	84.482
1	0	16	0.043	17.093	377.622	8719.894	93.380

Continúa en la siguiente página

w	r	hora	p	n	media	varianza	Desv. Est.
1	0	17	0.031	13.665	423.244	13532.258	116.328
1	0	18	0.028	12.204	424.280	15174.470	123.185
1	0	19	0.027	13.120	468.537	17200.322	131.150
1	0	20	0.023	11.064	475.805	20938.035	144.700
1	0	21	0.032	13.653	415.573	13064.586	114.300
1	0	22	0.033	11.061	320.537	9609.419	98.028
1	0	23	0.036	10.541	280.561	7748.319	88.025
1	1	0	0.023	6.339	271.250	11879.045	108.991
1	1	1	0.030	8.569	272.429	8934.031	94.520
1	1	2	0.042	10.541	238.143	5618.122	74.954
1	1	3	0.048	8.848	177.357	3732.444	61.094
1	1	4	0.061	7.923	122.036	2001.749	44.741
1	1	5	0.075	8.044	99.321	1325.647	36.409
1	1	6	0.104	9.438	81.036	776.820	27.871
1	1	7	0.079	4.067	47.643	605.801	24.613
1	1	8	0.052	4.099	74.607	1432.667	37.851
1	1	9	0.023	3.155	131.107	5579.667	74.697
1	1	10	0.023	4.368	184.536	7980.749	89.335
1	1	11	0.018	4.518	240.357	13028.015	114.140
1	1	12	0.016	4.630	294.000	18962.071	137.703
1	1	13	0.014	4.968	350.750	25112.616	158.470
1	1	14	0.013	4.576	360.821	28813.004	169.744
1	1	15	0.016	4.331	266.607	16678.453	129.145
1	1	16	0.015	4.652	297.964	19382.677	139.222
1	1	17	0.013	4.160	321.250	25127.902	158.518
1	1	18	0.012	3.790	311.536	25919.963	160.997
1	1	19	0.014	5.063	366.214	26853.668	163.871
1	1	20	0.014	4.628	335.393	24643.239	156.982
1	1	21	0.014	4.140	290.464	20668.820	143.767
1	1	22	0.026	5.699	215.429	8358.459	91.425
1	1	23	0.023	4.437	185.393	7932.239	89.063

Tabla G.1: Parámetros de la distribución Negbin por hora según tipo de día y lluvia. La columna w corresponde a la variable `weekend_holiday_flag` y la r a la variable `rain_flag`. Las columnas p y n son los parámetros de la distribución NegBin.

Apéndice H

Validación interna de la distribución Binomial Negativa con viajes del 2018

La Tabla [H.1](#) muestra el cálculo de los intervalos de confianza y estadísticas relevantes para el conjunto de datos sintético generados con las distribuciones Binomiales Negativas, a nivel horario y sujeto a diferentes condiciones (fines de semana/días festivos y si llueve o no).

hour	weekend_holiday_flag	rain_flag	percentage	numhi	numlow	numtotal	rangehi	rangelow
0	0	0	96.532	2	4	173	538	100
0	0	1	94.203	2	2	69	461	66
0	1	0	94.286	1	3	70	586	173
0	1	1	92.105	0	3	38	523	101
1	0	0	94.152	7	3	171	326	53
1	0	1	94.203	2	2	69	266	39
1	1	0	93.243	2	3	74	557	156
1	1	1	92.105	0	3	38	486	119
2	0	0	96.491	3	3	171	204	27
2	0	1	97.101	1	1	69	169	18
2	1	0	95.890	0	3	73	499	137
2	1	1	89.474	1	3	38	406	114
3	0	0	98.246	3	0	171	133	13
3	0	1	95.652	2	1	69	110	10
3	1	0	93.243	2	3	74	364	99
3	1	1	89.474	1	3	38	315	78
4	0	0	97.076	4	1	171	93	9
4	0	1	95.652	2	1	69	76	7
4	1	0	95.946	1	2	74	253	72
4	1	1	92.105	1	2	38	224	50
5	0	0	96.491	5	1	171	87	17
5	0	1	95.652	2	1	69	72	16
5	1	0	90.541	2	5	74	188	61
5	1	1	94.737	0	2	38	182	41
6	0	0	95.930	1	6	172	133	47
6	0	1	95.652	0	3	69	125	36
6	1	0	94.595	2	2	74	153	53
6	1	1	92.105	1	2	38	144	35
7	0	0	94.767	0	9	172	635	173
7	0	1	95.652	1	2	69	538	119
7	1	0	94.595	2	2	74	105	32

Continúa en la siguiente página

hour	weekend_holiday_flag	rain_flag	percentage	numhi	numlow	numtotal	rangehi	rangelow
7	1	1	100.000	0	0	38	106	12
8	0	0	93.023	0	12	172	1383	461
8	0	1	92.754	0	5	69	1335	307
8	1	0	93.243	2	3	74	167	42
8	1	1	94.737	0	2	38	165	19
9	0	0	93.605	0	11	172	1153	439
9	0	1	91.304	0	6	69	1202	248
9	1	0	95.946	2	1	74	301	58
9	1	1	92.105	0	3	38	313	27
10	0	0	95.930	2	5	172	565	240
10	0	1	92.754	0	5	69	585	96
10	1	0	94.595	2	2	74	387	105
10	1	1	92.105	0	3	38	396	53
11	0	0	97.661	2	2	171	492	230
11	0	1	88.406	0	8	69	528	74
11	1	0	93.243	2	3	74	475	163
11	1	1	89.474	0	4	38	510	71
12	0	0	95.906	6	1	171	528	250
12	0	1	88.571	0	8	70	592	80
12	1	0	94.595	1	3	74	563	197
12	1	1	89.474	0	4	38	619	89
13	0	0	95.906	6	1	171	686	323
13	0	1	91.304	0	6	69	776	113
13	1	0	93.243	2	3	74	683	237
13	1	1	94.737	0	2	38	722	112
14	0	0	95.322	4	4	171	1086	481
14	0	1	89.855	0	7	69	1188	186
14	1	0	97.297	0	2	74	691	247
14	1	1	92.105	0	3	38	761	108
15	0	0	94.152	4	6	171	1208	449
15	0	1	91.429	1	5	70	1182	168

Continúa en la siguiente página

hour	weekend_holiday_flag	rain_flag	percentage	numhi	numlow	numtotal	rangehi	rangelow
15	1	0	90.541	3	4	74	521	192
15	1	1	94.737	0	2	38	573	76
16	0	0	97.076	1	4	171	914	399
16	0	1	94.203	0	4	69	931	153
16	1	0	95.946	1	2	74	581	217
16	1	1	92.105	0	3	38	626	90
17	0	0	96.491	3	3	171	963	412
17	0	1	92.857	0	5	70	1014	170
17	1	0	94.595	2	2	74	679	226
17	1	1	92.105	0	3	38	698	89
18	0	0	97.076	0	5	171	1415	447
18	0	1	92.857	0	5	70	1448	176
18	1	0	91.892	4	2	74	697	218
18	1	1	92.105	0	3	38	697	80
19	0	0	98.246	0	3	171	1416	456
19	0	1	91.304	0	6	69	1380	164
19	1	0	93.243	3	2	74	758	247
19	1	1	92.105	0	3	38	750	118
20	0	0	97.076	3	2	171	1193	409
20	0	1	91.429	0	6	70	1115	156
20	1	0	91.892	4	2	74	798	235
20	1	1	94.737	0	2	38	705	101
21	0	0	98.246	0	3	171	906	334
21	0	1	95.652	0	3	69	823	134
21	1	0	97.297	0	2	74	667	222
21	1	1	97.368	1	0	38	632	80
22	0	0	97.076	2	3	171	662	201
22	0	1	94.286	0	4	70	582	94
22	1	0	94.595	2	2	74	539	158
22	1	1	94.737	1	1	38	428	75
23	0	0	97.076	1	4	171	587	154

Continúa en la siguiente página

hour	weekend_holiday_flag	rain_flag	percentage	numhi	numlow	numtotal	rangehi	rangelow
23	0	1	94.203	0	4	69	453	80
23	1	0	95.946	2	1	74	477	135
23	1	1	94.737	1	1	38	396	53

Tabla H.1: Cálculo de intervalos de confianza y estadísticas clave para datos sintéticos generados con distribuciones Binomiales Negativas, diferenciados por hora, tipo de día (laboral/festivo) y condiciones de lluvia.

Apéndice I

Función de Popularidad

Este apéndice incluye un análisis visual de la popularidad de las estaciones de BiciMad en 2018. En la Figura I.1, cada estación se representa con un punto en el mapa, cuyo color, basado en un gradiente, indica su porcentaje de popularidad como estación de origen, mientras que el tamaño del punto refleja su popularidad como estación de destino. Es importante observar que algunas estaciones presentan colores intensos en el gradiente y puntos de gran tamaño. Esto indica que son nodos altamente populares, tanto como estaciones de origen para tomar bicicletas y comenzar viajes, como estaciones de destino para dejarlas y finalizar trayectos. Además, la Tabla I.1 presenta, para cada identificador de estación, su latitud, longitud y los valores porcentuales de popularidad, tanto como estación de origen como de destino.

Popularidad por Estación 2018

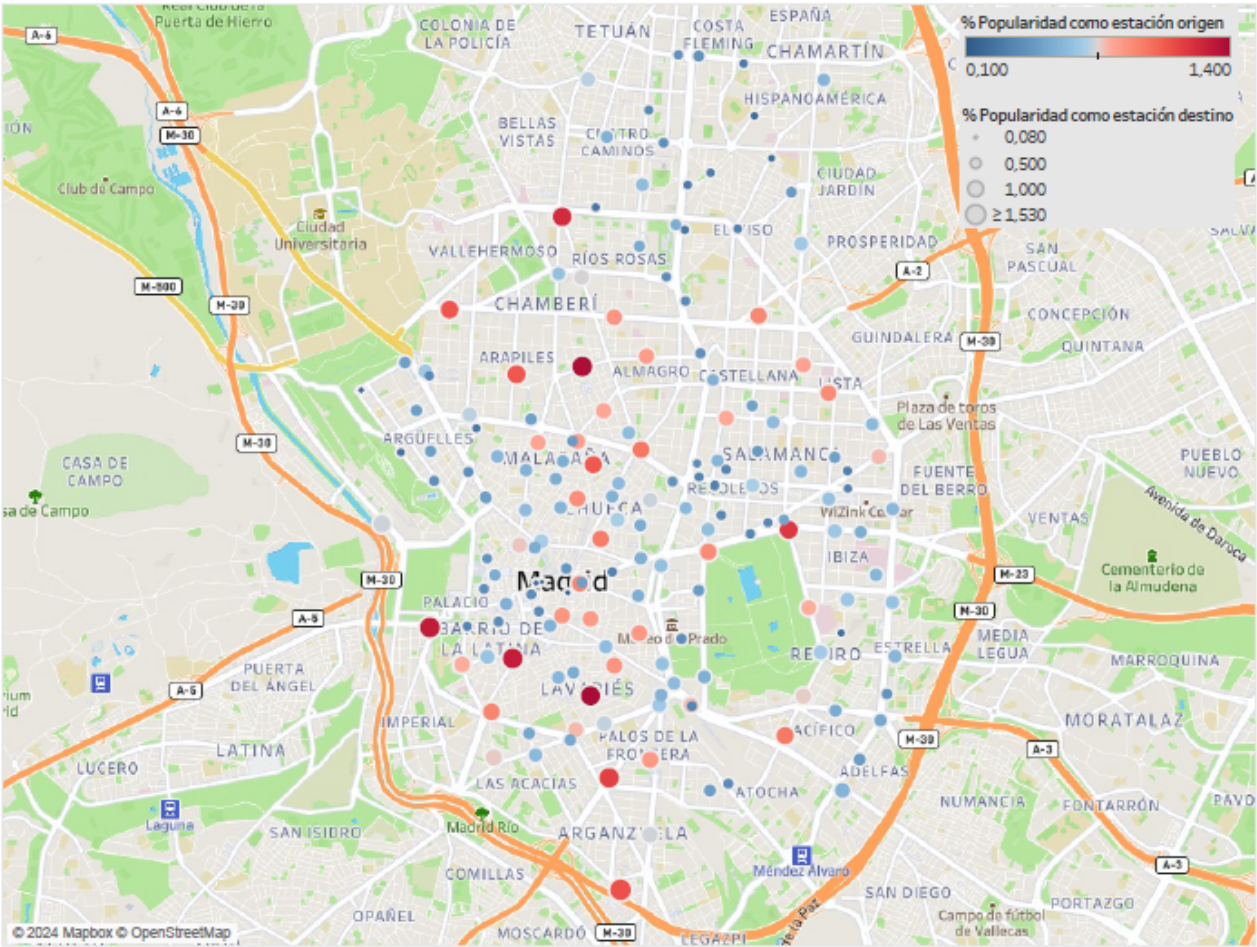


Figura I.1: Análisis visual de la popularidad de las estaciones de BiciMad en 2018. El color, indica su porcentaje de popularidad como estación de origen. El tamaño del punto refleja su popularidad como estación de destino. Fuente: Datos de viajes de BiciMad 2018 y visualización de elaboración propia.

ID Estación	Latitud	Longitud	Probabilidad origen	Probabilidad destino
1	40.4172137	-3.7018341	0.943	0.875
2	40.41731271	-3.701602938	0.456	0.402
3	40.4205886	-3.7058415	0.679	0.676
4	40.4302937	-3.7069171	0.466	0.468

Continúa en la siguiente página

ID Estación	Latitud	Longitud	Probabilidad origen	Probabilidad destino
5	40.4285524	-3.7025875	0.433	0.410
6	40.42852	-3.70205	0.857	0.819
7	40.424148	-3.698447	0.542	0.539
8	40.4251906	-3.6977715	0.588	0.566
9	40.4278682	-3.6954403	0.993	0.986
10	40.4156057	-3.7095084	0.538	0.542
11	40.42533	-3.69214	0.391	0.365
12	40.4269483	-3.7035918	0.523	0.512
13	40.4284246	-3.7061931	0.834	0.814
14	40.4273264	-3.7104417	0.570	0.560
15	40.4260957	-3.713479	0.373	0.363
16	40.4262383	-3.7074453	0.545	0.541
17	40.4230721	-3.7075065	0.595	0.580
18	40.4232649	-3.7038312	0.562	0.567
19	40.4207773	-3.6996502	0.980	0.975
20	40.4218616	-3.6954983	0.513	0.500
21	40.4192342	-3.6954615	0.533	0.548
25	40.4200783	-3.7065376	0.515	0.511
26	40.4202834	-3.7081246	0.772	0.763
27	40.417908	-3.710692	0.518	0.454
28	40.4173703	-3.7057791	0.128	0.080
29	40.4172781	-3.7063837	0.177	0.131
30	40.423855	-3.694475	0.723	0.694
31	40.4132798	-3.6956178	0.880	0.881
32	40.4181518	-3.6984368	0.311	0.318
33	40.4162619	-3.6957355	0.524	0.529
34	40.4192095	-3.711504	0.398	0.335
35	40.4163638	-3.7068969	0.368	0.362
36	40.4150099	-3.7061032	0.357	0.353
37	40.4165039	-3.7030833	0.114	0.114
38	40.4146755	-3.7036825	0.885	0.895
39	40.4141931	-3.7103285	0.366	0.360
40	40.4168	-3.71183	0.369	0.351
41	40.4108	-3.714113	0.807	0.871
42	40.4114041	-3.7114762	0.636	0.676
43	40.4112744	-3.7088337	1.302	1.418
44	40.4138846	-3.7049407	0.541	0.553
45	40.4122047	-3.6991147	0.605	0.615
46	40.4107085	-3.6982318	0.896	0.910
47	40.4101564	-3.7025024	0.516	0.509
48	40.4097617	-3.7040666	0.569	0.569
49	40.4070358	-3.7110513	0.948	0.994
50	40.4053153	-3.7071259	0.489	0.493
51	40.4047851	-3.7028265	0.498	0.474
52	40.4056107	-3.7022591	0.789	0.787
53	40.4060941	-3.6992759	0.718	0.748
54	40.4083684	-3.6933463	0.545	0.558
55	40.4075606	-3.6935205	0.660	0.666
56	40.4144226	-3.7007164	0.870	0.878
57	40.4083061	-3.7007111	1.395	1.337
58	40.4266828	-3.700423	1.083	1.055
59	40.4239757	-3.7020842	0.914	0.919
60	40.4184192	-3.7032414	0.412	0.405
61	40.4295658	-3.7126299	0.260	0.236

Continúa en la siguiente página

ID Estación	Latitud	Longitud	Probabilidad origen	Probabilidad destino
62	40.4222862	-3.697895	0.687	0.682
63	40.4255495	-3.7043418	0.525	0.527
64	40.419752	-3.688398	0.920	0.964
65	40.4213983	-3.6752045	0.653	0.664
66	40.4213148	-3.6724968	0.526	0.526
67	40.4179237	-3.6708959	0.568	0.560
69	40.4166834	-3.6894193	0.436	0.446
71	40.4108472	-3.693225	0.622	0.623
72	40.4128372	-3.6912023	0.369	0.386
73	40.409808	-3.688822	0.533	0.599
74	40.4153053	-3.6779232	0.809	0.781
75	40.4159569	-3.6738865	0.648	0.670
76	40.4157413	-3.6691838	0.629	0.630
77	40.4114475	-3.6689089	0.564	0.607
78	40.4117627	-3.6766813	0.670	0.671
79	40.4082805	-3.6784838	0.761	0.807
80	40.4071304	-3.6751352	0.510	0.525
81	40.40635	-3.670422	0.498	0.532
82	40.400789	-3.6744	0.556	0.781
83	40.4051451	-3.6803874	0.987	1.060
84	40.4075685	-3.6902255	0.775	0.798
85	40.4074902	-3.6901234	0.322	0.348
86	40.409297	-3.691987	0.595	0.583
87	40.4084556	-3.6697526	0.397	0.400
88	40.4132995	-3.6745323	0.204	0.173
89	40.4032501	-3.6726019	0.442	0.503
90	40.421501	-3.680008	1.197	1.161
91	40.4186516	-3.6933498	0.616	0.598
92	40.427736	-3.6832566	0.428	0.425
93	40.4313576	-3.6838303	0.466	0.433
94	40.4272582	-3.6752024	0.460	0.451
95	40.4250361	-3.6837876	0.688	0.630
96	40.426134	-3.6787441	0.571	0.579
97	40.4262945	-3.6865463	0.323	0.325
98	40.4257046	-3.6893698	0.312	0.287
99	40.4232153	-3.690756	0.579	0.520
100	40.4226584	-3.6870548	0.336	0.337
101	40.422064	-3.6821793	0.327	0.283
102	40.4222969	-3.6805189	0.481	0.385
103	40.4239447	-3.6758383	0.623	0.601
104	40.4261851	-3.6738714	0.297	0.285
105	40.42478	-3.67384	0.300	0.247
106	40.4231526	-3.6691524	0.575	0.579
107	40.4211802	-3.6840229	0.328	0.323
108	40.43037	-3.68653	0.821	0.807
109	40.42677	-3.68958	0.350	0.243
110	40.42699	-3.68745	0.565	0.528
111	40.4251002	-3.6877227	0.297	0.219
112	40.424963	-3.687745	0.348	0.279
113	40.4215246	-3.6884369	0.615	0.613
114	40.4273005	-3.6706024	0.786	0.784
115	40.4300481	-3.6816402	0.596	0.573
116	40.4347895	-3.7200845	0.458	0.463
117	40.4337322	-3.7175435	0.338	0.303

Continúa en la siguiente página

ID Estación	Latitud	Longitud	Probabilidad origen	Probabilidad destino
118	40.4341006	-3.7179687	0.657	0.638
119	40.4325991	-3.7246532	0.162	0.143
120	40.427657	-3.7205129	0.222	0.207
121	40.4277456	-3.7174158	0.451	0.454
122	40.4253944	-3.7170448	0.381	0.372
123	40.42402	-3.711603	0.465	0.409
124	40.42412	-3.711703	0.450	0.410
125	40.4309797	-3.7188898	0.445	0.454
126	40.400781	-3.6882407	0.452	0.505
127	40.4013216	-3.6863218	0.351	0.355
128	40.4032208	-3.6944768	0.876	0.943
129	40.4017926	-3.6987665	1.166	1.367
130	40.4291992	-3.6967169	0.602	0.591
131	40.4306458	-3.7133412	0.711	0.700
132	40.4219665	-3.7224983	0.729	1.018
133	40.4034076	-3.7108108	0.764	0.905
134	40.4035988	-3.7064516	0.585	0.623
135	40.3928821	-3.6975708	1.112	1.535
136	40.3972616	-3.6945025	0.732	0.827
137	40.4591366	-3.6894151	0.391	0.415
138	40.4585318	-3.684715	0.266	0.251
139	40.4572824	-3.7009675	0.706	0.695
140	40.4592351	-3.691533	0.384	0.318
141	40.4548456	-3.6946218	0.340	0.288
142	40.4527164	-3.6990077	0.535	0.515
143	40.4522454	-3.693085	0.369	0.347
144	40.4510188	-3.6817962	0.204	0.168
145	40.4489101	-3.6952943	0.599	0.568
146	40.4498613	-3.6881689	0.274	0.259
147	40.4488924	-3.6905604	0.266	0.255
148	40.4483269	-3.6797296	0.435	0.406
149	40.4463667	-3.7036675	1.245	1.248
150	40.447125	-3.7001669	0.246	0.239
151	40.4453307	-3.690861	0.255	0.245
152	40.445411	-3.6853312	0.298	0.267
153	40.4440297	-3.6956047	0.508	0.438
154	40.4457414	-3.6917932	0.520	0.406
155	40.4442258	-3.6787169	0.660	0.657
156	40.4418402	-3.7040344	0.621	0.574
157	40.44156	-3.70164	0.739	0.745
158	40.4415974	-3.6927821	0.361	0.305
159	40.4396792	-3.6907784	0.347	0.325
160	40.438994	-3.7154329	1.098	1.114
161	40.4383865	-3.6982353	0.890	0.872
162	40.4385127	-3.6831578	0.935	0.951
163	40.4344731	-3.7015686	1.388	1.408
164	40.435285	-3.6948626	0.862	0.836
165	40.4355143	-3.6892368	0.383	0.376
166	40.4345973	-3.678492	0.841	0.817
167	40.4334087	-3.6879154	0.545	0.486
168	40.4338516	-3.708439	1.091	1.176
169	40.4309524	-3.6993465	0.830	0.832
170	40.4323655	-3.6758555	0.922	0.907
171	40.429887	-3.6712823	0.582	0.563

Continúa en la siguiente página

ID Estación	Latitud	Longitud	Probabilidad origen	Probabilidad destino
172	40.4572466	-3.6763439	0.560	0.556
173	40.4543852	-3.6835926	0.200	0.192
174	40.4138333	-3.7135833	0.319	0.295
175	40.413736	-3.717487	1.314	1.438

Tabla I.1: Datos de popularidad

Apéndice J

**Código RStan empleado para la modelización
con enfoque bayesiano de las distancias**

En este apéndice se incluyen los código de R usando rstan para explorar la posibilidad de obtener los parámetros de las distribuciones Gamma ajustadas a los datos empíricos de distancias de los viajes de BiciMad, con una granularidad superior a la conseguida mediante el enfoque frecuentista.

Se comienza con un modelo bayesiano jerárquico (se explica al final por qué es jerárquico) relativamente sencillo, ya que se limita la estimación únicamente 7 grupos correspondientes a los días de la semana, en comparación con el modelo frecuentista presentado en la sección 4.2.2, el cual requería estimar 24 valores diferentes, uno para cada hora del día.

En este primer código, se buscan los parámetros de las distribuciones Gamma que modelan las distancias de los viajes de BiciMad agrupadas en cada uno de los 7 días de la semana. Mediante rstan, se define un modelo en lenguaje Stan. Cada día de la semana se considera un grupo independiente, para el cual se estiman los dos parámetros que definen la distribución Gamma: la forma (*shape*) y el inverso del *Scale* (*rate*). Los parámetros cada grupo son tratados como variables aleatorias que tienen como prior una distribución Normal. A su vez los parámetros media y desviación están definidos por los hiperparámetros *sigma* y *kappa*, respectivamente. Estos hiperparámetros, a su vez, tienen asignadas distribuciones prior exponenciales, lo que introduce una fuerte regularización hacia valores pequeños. Estas distribuciones previas reflejan incertidumbre inicial sobre los valores de los parámetros. El proceso de ajuste se realiza mediante muestreo con el algoritmo “*Monte Carlo Markov Chain*” (MCMC), implementado en la función `sampling` de rstan. Este enfoque genera muestras de las distribuciones posteriores de los parámetros, permitiendo una caracterización completa de la incertidumbre asociada a cada estimación. A continuación se muestra el código.

```

1
2 # MODELO CON GRANULARIDAD DIA DE LA SEMANA
3
4 library(rstan)
5 library(tictoc)
6 library(readr)
7
8 cores <- parallel::detectCores()-1
9 options(mc.cores = cores)
10
11 gammjer = '
12 data {
13   int<lower=2> K;
14   int n;
15   int group[n];
16   real dist[n];
17 }
18 parameters {
19   real<lower=0> shape[K];
20   real<lower=0> rate[K];
21   real<lower=0> sigma;
22   real<lower=0> kappa;
23 }
24
25
26 model {
27   dist ~ gamma(shape[group], rate[group]);
28   shape ~ normal(sigma, 3);
29   rate ~ normal(kappa, 3);
30   sigma ~ exponential(1);
31   kappa ~ exponential(1);
32 }
33 '
34
35 tic()
36 modelol=stan_model(model_code=gammjer)
37 toc()
38
39 # dataset18 <- read.csv('dataset18.csv', sep=";")
40 # saveRDS(dataset18, 'dataset18.rda')
41 #dataset18 <- readRDS('dataset18.rda')

```

```

42 n = 3249414
43 #index <- sample.int(3249414,n)
44 #data=list(K=7,n=n,dist=dataset18[index,"distance"],group=1+dataset18[index,"day_num"])
45
46 data=list(K=7,n=n,dist=dataset18[, "distance"],group=1+dataset18[, "day_num"])
47
48
49 tic()
50 fit = sampling(modelo1,data,cores = 1)
51 toc()
52 print(fit)
53 output = extract(fit)
54 # Compare posteriors between groups
55 par(mfrow=c(2,2))
56 plot(density(output$shape[,1]),xlim=c(2.,4),main=expression(alpha))
57 lines(density(output$shape[,2]),col=2)
58 plot(density(1/output$rate[,1]),xlim=c(500,1500),expression(1/beta))
59 lines(density(1/output$rate[,2]),col=2)
60 hist(output$sigma,main=expression(sigma),probability = TRUE)
61 hist(output$kappa,main=expression(kappa),probability = TRUE)
62
63
64
65
66 par(mfrow=c(2,2))
67 plot(density(output$shape[,1]),xlim=c(2.8,2.95),main=expression(alpha))
68 lines(density(output$shape[,2]),col=2)
69 lines(density(output$shape[,3]),col=3)
70 lines(density(output$shape[,4]),col=4)
71 lines(density(output$shape[,5]),col=5)
72 lines(density(output$shape[,6]),col=6)
73 lines(density(output$shape[,7]),col=7)
74 legend("topright", legend=c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "
    Sunday"),lty=1,col=1:7, cex=.4)
75
76
77 plot(density(1/output$rate[,1]),xlim=c(1000,1150),expression(1/beta))
78 lines(density(1/output$rate[,2]),col=2)
79 lines(density(1/output$rate[,3]),col=3)
80 lines(density(1/output$rate[,4]),col=4)
81 lines(density(1/output$rate[,5]),col=5)
82 lines(density(1/output$rate[,6]),col=6)
83 lines(density(1/output$rate[,7]),col=7)
84 legend("topright", legend=c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "
    Sunday"),lty=1,col=1:7, cex=.4)
85
86 hist(output$sigma,main=expression(sigma),probability = TRUE)
87 hist(output$kappa,main=expression(kappa),probability = TRUE)
88
89 saveRDS(fit, 'fitdaysallv4.rda')
90 saveRDS(output, 'outputall12018days.rda')

```

La Figura J.1 muestra el resultado visual obtenido tras la ejecución del código.

Como se puede apreciar en la salida por consola de la ejecución, este proceso tardó 143003.67 segundos que son aproximadamente 40 horas en un ordenador de sobremesa con procesador Intel Core i7 (con 4 núcleos de los cuales por configuración del propio script se utilizaban todos menos 1) y 16GB de RAM.

```

> tic()
> fit = sampling(modelo1,data,cores = 1)

SAMPLING FOR MODEL 'e92fb283fa4edb6119e6764cae9f39c9' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0.58 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 5800 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:

```

```

Chain 1: Iteration:    1 / 2000 [  0%] (Warmup)
Chain 1: Iteration:   200 / 2000 [ 10%] (Warmup)
Chain 1: Iteration:   400 / 2000 [ 20%] (Warmup)
Chain 1: Iteration:   600 / 2000 [ 30%] (Warmup)
Chain 1: Iteration:   800 / 2000 [ 40%] (Warmup)
Chain 1: Iteration:  1000 / 2000 [ 50%] (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%] (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%] (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%] (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%] (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%] (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%] (Sampling)
Chain 1:
Chain 1: Elapsed Time: 21844.5 seconds (Warm-up)
Chain 1:           12684.5 seconds (Sampling)
Chain 1:           34529 seconds (Total)
Chain 1:

SAMPLING FOR MODEL 'e92fb283fa4edb6119e6764cae9f39c9' NOW (CHAIN 2).
Chain 2:
Chain 2: Gradient evaluation took 0.578 seconds
Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 5780 seconds.
Chain 2: Adjust your expectations accordingly!
Chain 2:
Chain 2:
Chain 2: Iteration:    1 / 2000 [  0%] (Warmup)
Chain 2: Iteration:   200 / 2000 [ 10%] (Warmup)
Chain 2: Iteration:   400 / 2000 [ 20%] (Warmup)
Chain 2: Iteration:   600 / 2000 [ 30%] (Warmup)
Chain 2: Iteration:   800 / 2000 [ 40%] (Warmup)
Chain 2: Iteration:  1000 / 2000 [ 50%] (Warmup)
Chain 2: Iteration: 1001 / 2000 [ 50%] (Sampling)
Chain 2: Iteration: 1200 / 2000 [ 60%] (Sampling)
Chain 2: Iteration: 1400 / 2000 [ 70%] (Sampling)
Chain 2: Iteration: 1600 / 2000 [ 80%] (Sampling)
Chain 2: Iteration: 1800 / 2000 [ 90%] (Sampling)
Chain 2: Iteration: 2000 / 2000 [100%] (Sampling)
Chain 2:
Chain 2: Elapsed Time: 22823.9 seconds (Warm-up)
Chain 2:           13150.5 seconds (Sampling)
Chain 2:           35974.4 seconds (Total)
Chain 2:

SAMPLING FOR MODEL 'e92fb283fa4edb6119e6764cae9f39c9' NOW (CHAIN 3).
Chain 3:
Chain 3: Gradient evaluation took 0.563 seconds
Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 5630 seconds.
Chain 3: Adjust your expectations accordingly!
Chain 3:
Chain 3:
Chain 3: Iteration:    1 / 2000 [  0%] (Warmup)
Chain 3: Iteration:   200 / 2000 [ 10%] (Warmup)
Chain 3: Iteration:   400 / 2000 [ 20%] (Warmup)
Chain 3: Iteration:   600 / 2000 [ 30%] (Warmup)
Chain 3: Iteration:   800 / 2000 [ 40%] (Warmup)
Chain 3: Iteration:  1000 / 2000 [ 50%] (Warmup)
Chain 3: Iteration: 1001 / 2000 [ 50%] (Sampling)
Chain 3: Iteration: 1200 / 2000 [ 60%] (Sampling)
Chain 3: Iteration: 1400 / 2000 [ 70%] (Sampling)
Chain 3: Iteration: 1600 / 2000 [ 80%] (Sampling)
Chain 3: Iteration: 1800 / 2000 [ 90%] (Sampling)
Chain 3: Iteration: 2000 / 2000 [100%] (Sampling)
Chain 3:
Chain 3: Elapsed Time: 21620 seconds (Warm-up)
Chain 3:           12890.2 seconds (Sampling)
Chain 3:           34510.2 seconds (Total)
Chain 3:

SAMPLING FOR MODEL 'e92fb283fa4edb6119e6764cae9f39c9' NOW (CHAIN 4).
Chain 4:

```

```
Chain 4: Gradient evaluation took 0.5 seconds
Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 5000 seconds.
Chain 4: Adjust your expectations accordingly!
Chain 4:
Chain 4:
Chain 4: Iteration:    1 / 2000 [  0%] (Warmup)
Chain 4: Iteration:   200 / 2000 [ 10%] (Warmup)
Chain 4: Iteration:   400 / 2000 [ 20%] (Warmup)
Chain 4: Iteration:   600 / 2000 [ 30%] (Warmup)
Chain 4: Iteration:   800 / 2000 [ 40%] (Warmup)
Chain 4: Iteration:  1000 / 2000 [ 50%] (Warmup)
Chain 4: Iteration:  1001 / 2000 [ 50%] (Sampling)
Chain 4: Iteration:  1200 / 2000 [ 60%] (Sampling)
Chain 4: Iteration:  1400 / 2000 [ 70%] (Sampling)
Chain 4: Iteration:  1600 / 2000 [ 80%] (Sampling)
Chain 4: Iteration:  1800 / 2000 [ 90%] (Sampling)
Chain 4: Iteration:  2000 / 2000 [100%] (Sampling)
Chain 4:
Chain 4: Elapsed Time: 25095.5 seconds (Warm-up)
Chain 4:                12837.3 seconds (Sampling)
Chain 4:                37932.8 seconds (Total)
Chain 4:
> toc()
143003.67 sec elapsed
> print(fit)
Inference for Stan model: e92fb283fa4edb6119e6764cae9f39c9.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.
```

	mean	se_mean	sd	2.5%	25%
shape[1]	2.85	0.00	0.01	2.84	2.85
shape[2]	2.86	0.00	0.01	2.85	2.86
shape[3]	2.86	0.00	0.01	2.85	2.86
shape[4]	2.86	0.00	0.01	2.85	2.85
shape[5]	2.86	0.00	0.01	2.85	2.86
shape[6]	2.87	0.00	0.01	2.86	2.87
shape[7]	2.88	0.00	0.01	2.86	2.87
rate[1]	0.00	0.00	0.00	0.00	0.00
rate[2]	0.00	0.00	0.00	0.00	0.00
rate[3]	0.00	0.00	0.00	0.00	0.00
rate[4]	0.00	0.00	0.00	0.00	0.00
rate[5]	0.00	0.00	0.00	0.00	0.00
rate[6]	0.00	0.00	0.00	0.00	0.00
rate[7]	0.00	0.00	0.00	0.00	0.00
sigma	1.75	0.02	0.98	0.13	1.00
kappa	0.56	0.01	0.49	0.02	0.18
lp__	-28541312.86	0.08	3.05	-28541320.07	-28541314.61
	50%	75%	97.5%	n_eff	Rhat
shape[1]	2.85	2.85	2.86	5005	1
shape[2]	2.86	2.87	2.87	5344	1
shape[3]	2.86	2.87	2.87	4806	1
shape[4]	2.86	2.86	2.87	5278	1
shape[5]	2.86	2.86	2.87	5127	1
shape[6]	2.87	2.88	2.89	4330	1
shape[7]	2.88	2.88	2.89	3776	1
rate[1]	0.00	0.00	0.00	4992	1
rate[2]	0.00	0.00	0.00	5273	1
rate[3]	0.00	0.00	0.00	4917	1
rate[4]	0.00	0.00	0.00	5529	1
rate[5]	0.00	0.00	0.00	4966	1
rate[6]	0.00	0.00	0.00	4321	1
rate[7]	0.00	0.00	0.00	3922	1
sigma	1.69	2.40	3.83	1873	1
kappa	0.42	0.81	1.79	3800	1
lp__	-28541312.41	-28541310.72	-28541308.04	1468	1

Samples were drawn using NUTS(diag_e) at Tue Dec 07 14:23:18 2021.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1)

De forma análoga se facilita a continuación otros códigos pero con granularidad horaria, mensual y al final uno con granularidad semanal. Nótese

```

1
2 # MODELO CON GRANULARIDAD HORARIA
3
4 library(rstan)
5 library(tictoc)
6 library(readr)
7
8 cores <- parallel::detectCores()-1
9 options(mc.cores = cores)
10
11 gammjer = '
12 data {
13   int<lower=2> K;
14   int n;
15   int group[n];
16   real dist[n];
17 }
18 parameters {
19   real<lower=0> shape[K];
20   real<lower=0> rate[K];
21   real<lower=0> sigma;
22   real<lower=0> kappa;
23 }
24
25
26 model {
27   dist ~ gamma(shape[group],rate[group]);
28   shape ~ normal(sigma,3);
29   rate ~ normal(kappa,3);
30   sigma ~ exponential(1);
31   kappa ~ exponential(1);
32 }
33 '
34 tic()
35 modelo1=stan_model(model_code=gammjer)
36 toc()
37
38 # dataset18 <- read.csv('dataset18.csv',sep=";")
39 # saveRDS(dataset18,'dataset18.rda')
40 #dataset18 <- readRDS('dataset18.rda')
41 n = 3249414
42 #index <- sample.int(3249414,n)
43 data=list(K=24,n=n,dist=dataset18[, "distance"],group=1+dataset18[, "hour"])
44
45 tic()
46 fit = sampling(modelo1,data,cores = 1)
47 toc()
48 print(fit)
49 output = extract(fit)
50 # Compare posteriors between groups
51 #par(mfrow=c(2,2))
52 #plot(density(output$shape[,1]),xlim=c(2.,4),main=expression(alpha))
53 #lines(density(output$shape[,2]),col=2)
54 #plot(density(1/output$rate[,1]),xlim=c(500,1500),expression(1/beta))
55 #lines(density(1/output$rate[,2]),col=2)
56 #hist(output$sigma,main=expression(sigma),probability = TRUE)
57 #hist(output$kappa,main=expression(kappa),probability = TRUE)
58
59
60
61 par(mfrow=c(2,2))
62 plot(density(output$shape[,1]),xlim=c(2.69,3.25),main=expression(alpha))
63 lines(density(output$shape[,2]),col=2)
64 lines(density(output$shape[,3]),col=3)

```

```

65 lines(density(output$shape[,4]),col=4)
66 lines(density(output$shape[,5]),col=5)
67 lines(density(output$shape[,6]),col=6)
68 lines(density(output$shape[,7]),col=7)
69 lines(density(output$shape[,8]),col=8)
70 lines(density(output$shape[,9]),col=9)
71 lines(density(output$shape[,10]),col=10)
72 lines(density(output$shape[,11]),col=11)
73 lines(density(output$shape[,12]),col=12)
74 lines(density(output$shape[,13]),col=13)
75 lines(density(output$shape[,14]),col=14)
76 lines(density(output$shape[,15]),col=15)
77 lines(density(output$shape[,16]),col=16)
78 lines(density(output$shape[,17]),col=17)
79 lines(density(output$shape[,18]),col=18)
80 lines(density(output$shape[,19]),col=19)
81 lines(density(output$shape[,20]),col=20)
82 lines(density(output$shape[,21]),col=21)
83 lines(density(output$shape[,22]),col=22)
84 lines(density(output$shape[,23]),col=23)
85 lines(density(output$shape[,24]),col=24)
86 legend("topright", legend=c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12", "13", "14",
    "15", "16", "17", "18", "19", "20", "21", "22", "23", "24"),lty=1,col=1:24, cex=.30)
87
88
89 plot(density(1/output$rate[,1]),xlim=c(900,1150),expression(1/beta))
90 lines(density(1/output$rate[,2]),col=2)
91 lines(density(1/output$rate[,3]),col=3)
92 lines(density(1/output$rate[,4]),col=4)
93 lines(density(1/output$rate[,5]),col=5)
94 lines(density(1/output$rate[,6]),col=6)
95 lines(density(1/output$rate[,7]),col=7)
96 lines(density(1/output$rate[,8]),col=8)
97 lines(density(1/output$rate[,9]),col=9)
98 lines(density(1/output$rate[,10]),col=10)
99 lines(density(1/output$rate[,11]),col=11)
100 lines(density(1/output$rate[,12]),col=12)
101 lines(density(1/output$rate[,13]),col=13)
102 lines(density(1/output$rate[,14]),col=14)
103 lines(density(1/output$rate[,15]),col=15)
104 lines(density(1/output$rate[,16]),col=16)
105 lines(density(1/output$rate[,17]),col=17)
106 lines(density(1/output$rate[,18]),col=18)
107 lines(density(1/output$rate[,19]),col=19)
108 lines(density(1/output$rate[,20]),col=20)
109 lines(density(1/output$rate[,21]),col=21)
110 lines(density(1/output$rate[,22]),col=22)
111 lines(density(1/output$rate[,23]),col=23)
112 lines(density(1/output$rate[,24]),col=24)
113 legend("topright", legend=c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12", "13", "14",
    "15", "16", "17", "18", "19", "20", "21", "22", "23", "24"),lty=1,col=1:24, cex=.30)
114
115 hist(output$sigma,main=expression(sigma),probability = TRUE)
116 hist(output$kappa,main=expression(kappa),probability = TRUE)
117
118 saveRDS(fit, 'fithoursall2018.rda')
119 saveRDS(output, 'outputall2018hours.rda')

```

```

1
2 # MODELO CON GRANULARIDAD MENSUAL
3 library(rstan)
4 library(tictoc)
5 library(readr)
6
7 cores <- parallel::detectCores()-1
8 options(mc.cores = cores)
9
10 gammjer = '
11 data {
12   int<lower=2> K;

```

```

13   int n;
14   int group[n];
15   real dist[n];
16 }
17 parameters {
18   real<lower=0> shape[K];
19   real<lower=0> rate[K];
20   real<lower=0> sigma;
21   real<lower=0> kappa;
22 }
23
24
25 model {
26   dist ~ gamma(shape[group],rate[group]);
27   shape ~ normal(sigma,3);
28   rate ~ normal(kappa,3);
29   sigma ~ exponential(1);
30   kappa ~ exponential(1);
31 }
32 '
33 tic()
34 modelol=stan_model(model_code=gammajer)
35 toc()
36
37 # dataset18 <- read.csv('dataset18.csv',sep=";")
38 # saveRDS(dataset18,'dataset18.rda')
39 #dataset18 <- readRDS('dataset18.rda')
40 n = 3249414
41 #index <- sample.int(3249414,n)
42 data=list(K=12,n=n,dist=dataset18[, "distance"],group=dataset18[, "month_num"])
43
44 tic()
45 fit = sampling(modelol,data,cores = 1)
46 toc()
47 print(fit)
48 output = extract(fit)
49 # Compare posteriors between groups
50 #par(mfrow=c(2,2))
51 #plot(density(output$shape[,1]),xlim=c(2.,4),main=expression(alpha))
52 #lines(density(output$shape[,2]),col=2)
53 #plot(density(1/output$rate[,1]),xlim=c(500,1500),expression(1/beta))
54 #lines(density(1/output$rate[,2]),col=2)
55 #hist(output$sigma,main=expression(sigma),probability = TRUE)
56 #hist(output$kappa,main=expression(kappa),probability = TRUE)
57
58
59
60 par(mfrow=c(2,2))
61 plot(density(output$shape[,1]),xlim=c(2.75,3.0),main=expression(alpha))
62 lines(density(output$shape[,2]),col=2)
63 lines(density(output$shape[,3]),col=3)
64 lines(density(output$shape[,4]),col=4)
65 lines(density(output$shape[,5]),col=5)
66 lines(density(output$shape[,6]),col=6)
67 lines(density(output$shape[,7]),col=7)
68 lines(density(output$shape[,8]),col=8)
69 lines(density(output$shape[,9]),col=9)
70 lines(density(output$shape[,10]),col=10)
71 lines(density(output$shape[,11]),col=11)
72 lines(density(output$shape[,12]),col=12)
73 legend("topright", legend=c("Jan", "Feb", "Mar", "Apr", "May", "June", "July", "Aug", "Sep", "Oct", "
    Nov", "Dec"),lty=1,col=1:12, cex=.30)
74
75
76 plot(density(1/output$rate[,1]),xlim=c(1000,1100),expression(1/beta))
77 lines(density(1/output$rate[,2]),col=2)
78 lines(density(1/output$rate[,3]),col=3)
79 lines(density(1/output$rate[,4]),col=4)
80 lines(density(1/output$rate[,5]),col=5)
81 lines(density(1/output$rate[,6]),col=6)

```

```

82 lines(density(1/output$rate[,7]),col=7)
83 lines(density(1/output$rate[,8]),col=8)
84 lines(density(1/output$rate[,9]),col=9)
85 lines(density(1/output$rate[,10]),col=10)
86 lines(density(1/output$rate[,11]),col=11)
87 lines(density(1/output$rate[,12]),col=12)
88 legend("topright", legend=c("Jan", "Feb", "Mar", "Apr", "May", "June", "July", "Aug", "Sep", "Oct", "
    Nov", "Dec"),lty=1,col=1:12, cex=.30)
89
90 hist(output$sigma,main=expression(sigma),probability = TRUE)
91 hist(output$kappa,main=expression(kappa),probability = TRUE)
92
93 saveRDS(fit,'fitmonthsall2018.rda')
94 saveRDS(output,'outputall2018months.rda')
95 #plot (fit)

```

```

1
2 # MODELO CON GRANULARIDAD NUMERO DE SEMANA
3
4 library(rstan)
5 library(tictoc)
6
7
8 cores <- parallel::detectCores()-1
9 options(mc.cores = cores)
10
11 gammjer = '
12 data {
13   int<lower=2> K;
14   int n;
15   int group[n];
16   real dist[n];
17 }
18 parameters {
19   real<lower=0> shape[K];
20   real<lower=0> rate[K];
21   real<lower=0> sigma;
22   real<lower=0> kappa;
23 }
24
25
26 model {
27   dist ~ gamma(shape[group],rate[group]);
28   shape ~ normal(sigma,3);
29   rate ~ normal(kappa,3);
30   sigma ~ exponential(1);
31   kappa ~ exponential(1);
32 }
33 '
34 tic()
35 modelo1=stan_model(model_code=gammjer)
36 toc()
37
38 # dataset18 <- read.csv('dataset18.csv',sep=";")
39 # saveRDS(dataset18,'dataset18.rda')
40 #dataset18 <- readRDS('dataset18.rda')
41 n = 3249414
42 #index <- sample.int(3249414,n)
43 data=list(K=52,n=n,dist=dataset18[, "distance"],group=dataset18[, "week_num"])
44
45 tic()
46 fit = sampling(modelo1,data,cores = 1)
47 toc()
48 print(fit)
49 output = extract(fit)
50 # Compare posteriors between groups
51 #par(mfrow=c(2,2))
52 #plot(density(output$shape[,1]),xlim=c(2.,4),main=expression(alpha))
53 #lines(density(output$shape[,2]),col=2)
54 #plot(density(1/output$rate[,1]),xlim=c(500,1500),expression(1/beta))

```

```

55 #lines(density(1/output$rate[,2]),col=2)
56 #hist(output$sigma,main=expression(sigma),probability = TRUE)
57 #hist(output$kappa,main=expression(kappa),probability = TRUE)
58
59
60
61 par(mfrow=c(2,2))
62 plot(density(output$shape[,1]),xlim=c(2.7,3.05),main=expression(alpha))
63 lines(density(output$shape[,2]),col=2)
64 lines(density(output$shape[,3]),col=3)
65 lines(density(output$shape[,4]),col=4)
66 lines(density(output$shape[,5]),col=5)
67 lines(density(output$shape[,6]),col=6)
68 lines(density(output$shape[,7]),col=7)
69 lines(density(output$shape[,8]),col=8)
70 lines(density(output$shape[,9]),col=9)
71 lines(density(output$shape[,10]),col=10)
72 lines(density(output$shape[,11]),col=11)
73 lines(density(output$shape[,12]),col=12)
74 lines(density(output$shape[,13]),col=13)
75 lines(density(output$shape[,14]),col=14)
76 lines(density(output$shape[,15]),col=15)
77 lines(density(output$shape[,16]),col=16)
78 lines(density(output$shape[,17]),col=17)
79 lines(density(output$shape[,18]),col=18)
80 lines(density(output$shape[,19]),col=19)
81 lines(density(output$shape[,20]),col=20)
82 lines(density(output$shape[,21]),col=21)
83 lines(density(output$shape[,22]),col=22)
84 lines(density(output$shape[,23]),col=23)
85 lines(density(output$shape[,24]),col=24)
86 lines(density(output$shape[,25]),col=25)
87 lines(density(output$shape[,26]),col=26)
88 lines(density(output$shape[,27]),col=27)
89 lines(density(output$shape[,28]),col=28)
90 lines(density(output$shape[,29]),col=29)
91 lines(density(output$shape[,30]),col=30)
92 lines(density(output$shape[,31]),col=31)
93 lines(density(output$shape[,32]),col=32)
94 lines(density(output$shape[,33]),col=33)
95 lines(density(output$shape[,34]),col=34)
96 lines(density(output$shape[,35]),col=35)
97 lines(density(output$shape[,36]),col=36)
98 lines(density(output$shape[,37]),col=37)
99 lines(density(output$shape[,38]),col=38)
100 lines(density(output$shape[,39]),col=39)
101 lines(density(output$shape[,40]),col=40)
102 lines(density(output$shape[,41]),col=41)
103 lines(density(output$shape[,42]),col=42)
104 lines(density(output$shape[,43]),col=43)
105 lines(density(output$shape[,44]),col=44)
106 lines(density(output$shape[,45]),col=45)
107 lines(density(output$shape[,46]),col=46)
108 lines(density(output$shape[,47]),col=47)
109 lines(density(output$shape[,48]),col=48)
110 lines(density(output$shape[,49]),col=49)
111 lines(density(output$shape[,50]),col=50)
112 lines(density(output$shape[,51]),col=51)
113 lines(density(output$shape[,52]),col=52)
114
115 legend("topright", legend=c("1","2","3","4","5","6","7","8","9","10","11","12","13","14",
116 "15","16","17","18","19","20","21","22","23","24","25","26","27",
117 "28","29","30","31","32","33","34","35","36","37","38","39","40",
118 "41","42","43","44","45","46","47","48","49","50","51","52"),lty=1,col
    =1:52, cex=.30)
119
120
121 plot(density(1/output$rate[,1]),xlim=c(990,1130),expression(1/beta))
122 lines(density(1/output$rate[,2]),col=2)
123 lines(density(1/output$rate[,3]),col=3)

```

```

124 lines(density(1/output$rate[,4]),col=4)
125 lines(density(1/output$rate[,5]),col=5)
126 lines(density(1/output$rate[,6]),col=6)
127 lines(density(1/output$rate[,7]),col=7)
128 lines(density(1/output$rate[,8]),col=8)
129 lines(density(1/output$rate[,9]),col=9)
130 lines(density(1/output$rate[,10]),col=10)
131 lines(density(1/output$rate[,11]),col=11)
132 lines(density(1/output$rate[,12]),col=12)
133 lines(density(1/output$rate[,13]),col=13)
134 lines(density(1/output$rate[,14]),col=14)
135 lines(density(1/output$rate[,15]),col=15)
136 lines(density(1/output$rate[,16]),col=16)
137 lines(density(1/output$rate[,17]),col=17)
138 lines(density(1/output$rate[,18]),col=18)
139 lines(density(1/output$rate[,19]),col=19)
140 lines(density(1/output$rate[,20]),col=20)
141 lines(density(1/output$rate[,21]),col=21)
142 lines(density(1/output$rate[,22]),col=22)
143 lines(density(1/output$rate[,23]),col=23)
144 lines(density(1/output$rate[,24]),col=24)
145 lines(density(1/output$rate[,25]),col=25)
146 lines(density(1/output$rate[,26]),col=26)
147 lines(density(1/output$rate[,27]),col=27)
148 lines(density(1/output$rate[,28]),col=28)
149 lines(density(1/output$rate[,29]),col=29)
150 lines(density(1/output$rate[,30]),col=30)
151 lines(density(1/output$rate[,31]),col=31)
152 lines(density(1/output$rate[,32]),col=32)
153 lines(density(1/output$rate[,33]),col=33)
154 lines(density(1/output$rate[,34]),col=34)
155 lines(density(1/output$rate[,35]),col=35)
156 lines(density(1/output$rate[,36]),col=36)
157 lines(density(1/output$rate[,37]),col=37)
158 lines(density(1/output$rate[,38]),col=38)
159 lines(density(1/output$rate[,39]),col=39)
160 lines(density(1/output$rate[,40]),col=40)
161 lines(density(1/output$rate[,41]),col=41)
162 lines(density(1/output$rate[,42]),col=42)
163 lines(density(1/output$rate[,43]),col=43)
164 lines(density(1/output$rate[,44]),col=44)
165 lines(density(1/output$rate[,45]),col=45)
166 lines(density(1/output$rate[,46]),col=46)
167 lines(density(1/output$rate[,47]),col=47)
168 lines(density(1/output$rate[,48]),col=48)
169 lines(density(1/output$rate[,49]),col=49)
170 lines(density(1/output$rate[,50]),col=50)
171 lines(density(1/output$rate[,51]),col=51)
172 lines(density(1/output$rate[,52]),col=52)
173
174 legend("topright", legend=c("1","2","3","4","5","6","7","8","9","10","11","12","13","14",
175                             "15","16","17","18","19","20","21","22","23","24","25","26","27",
176                             "28","29","30","31","32","33","34","35","36","37","38","39","40",
177                             "41","42","43","44","45","46","47","48","49","50","51","52"),lty=1,col
                             =1:52, cex=.30)
178 hist(output$sigma,main=expression(sigma),probability = TRUE)
179 hist(output$kappa,main=expression(kappa),probability = TRUE)
180
181 saveRDS(fit,'fithweeknumall2018.rda')
182 saveRDS(output,'outputall2018weeknum.rda')

```

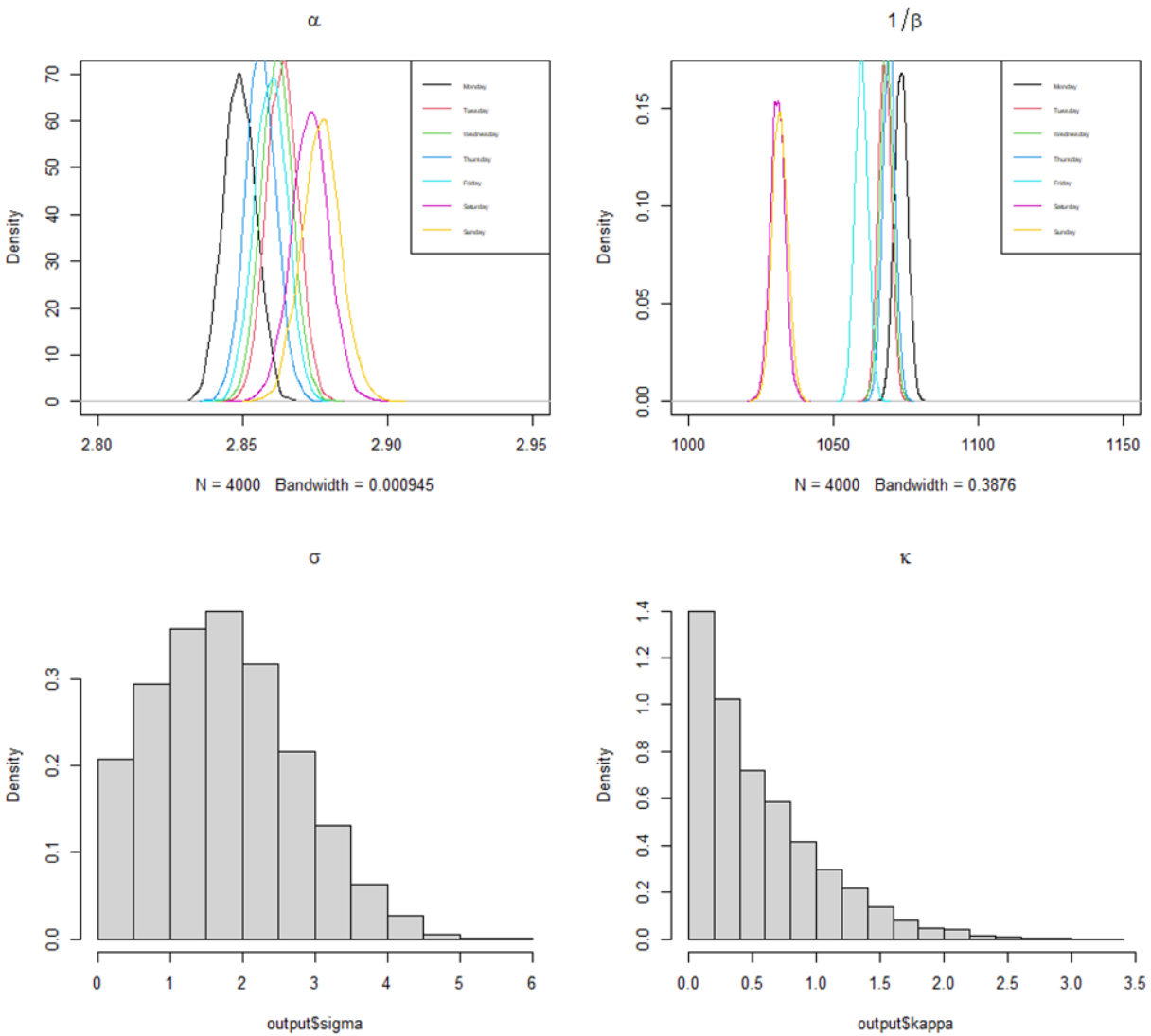


Figura J.1: Salida visual de la ejecución del Modelo bayesiano jerárquico para ajustar la distribución Gamma a grupos a nivel de día de la semana. Los colores corresponden a los días de la siguiente forma: negro al lunes, rosa al martes, verde al miércoles, azul al jueves, azul celeste al viernes, fucsia al sábado y finalmente amarillo al domingo.

Bibliografía

- Abramovich, F. and Ritov, Y. (2022). *Statistical theory: a concise introduction*. Chapman and Hall/CRC.
- AEMET [Agencia Estatal de Meteorología de España] (2024a). Open data api.
<https://opendata.aemet.es/dist/index.html?> (accedido en 13 Feb 2023).
- AEMET [Agencia Estatal de Meteorología de España] (2024b). Open data web.
https://www.aemet.es/es/datos_abiertos/AEMET_OpenData (accedido en 13 Feb 2023).
- Albuquerque, V., Dias, M. S., and Bacao, F. (2021). Machine learning approaches to bike-sharing systems: A systematic literature review. *ISPRS International Journal of Geo-Information*, 10(2).
- Alcorn, L. G. and Jiao, J. (2023). Bike-sharing station usage and the surrounding built environments in major texas cities. *Journal of planning education and research*, 43(1):122–135.
- Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.
- Alvarez-Valdes, R., Belenguer, J. M., Benavent, E., Bermudez, J. D., Muñoz, F., Vercher, E., and Verdejo, F. (2016). Optimizing the level of service quality of a bike-sharing system. *Omega (United Kingdom)*, 62:163–175.
- Andersson, A., Winslott Hiselius, L., and Adell, E. (2018). Promoting sustainable travel behaviour through the use of smartphone applications: A review and development of a conceptual model. *Travel Behaviour and Society*, 11(January):52–61.
- Applegate, D. L. (2006). *The traveling salesman problem: a computational study*, volume 17. Princeton university press.
- Arias-Molinares, D. and García-Palomares, J. C. (2020). The Ws of MaaS: Understanding mobility as a service from literature review. *IATSS Research*.
- Ayuntamiento de Madrid, Portal de Datos Abiertos (2024). Calendario laboral.
<https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnnextoid=9f710c96da3f9510VgnVCM2000001f4a900aRCRD&vgnnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnnextfmt=default> (accedido en 13 Feb 2023).

- Bakogiannis, E., Siti, M., Tsigdinos, S., Vassi, A., and Nikitas, A. (2019). Monitoring the first dockless bike sharing system in greece: Understanding user perceptions, usage patterns and adoption barriers. *Research in Transportation Business & Management*, 33:100432.
- Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.
- Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., and Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, 734:1–74.
- Barbour, N., Zhang, Y., and Mannering, F. (2019). A statistical analysis of bike sharing usage and its potential as an auto-trip substitute. *Journal of Transport & Health*, 12:253–262.
- Bazzani, A., Giorgini, B., Rambaldi, S., Gallotti, R., and Giovannini, L. (2010). Statistical laws in urban mobility from microscopic gps data in the area of florence. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(05):P05001.
- Beecham, R., Wood, J., and Bowerman, A. (2014). Studying commuting behaviours using collaborative visual analytics. *Computers, Environment and Urban Systems*, 47:5–15.
- Ben-Akiva, M. and Michel, B. (1999). Discrete Choice Methods and their Applications to Short Term Travel Decisions. In Hall, R. W., editor, *Handbook of Transportation Science*. Springer US.
- Benchimol, M., Benchimol, P., Chappert, B., De La Taille, A., Laroche, F., Meunier, F., and Robinet, L. (2011). Balancing the stations of a self service “bike hire” system. *RAIRO-Operations Research-Recherche Opérationnelle*, 45(1):37–61.
- Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.
- Bolstad, W. M. and Curran, J. M. (2016). *Introduction to Bayesian statistics*. John Wiley & Sons.
- Bordagaray, M., Dell’Olio, L., Fonzone, A., and Ibeas, Á. (2016). Capturing the conditions that introduce systematic variation in bike-sharing travel behavior using data mining techniques. *Transportation research part C: emerging technologies*, 71:231–248.
- Borgnat, P., Abry, P., Flandrin, P., Robardet, C., Rouquier, J. B., and Fleury, E. (2011). Shared bicycles in a city: A signal processing and data analysis perspective. *Advances in Complex Systems*, 14(3):415–438.
- Botsman, R. and Rogers, R. (2010). *What’s Mine Is Yours: The Rise of Collaborative Consumption*. HarperCollins.
- Bouveyron, C., Côme, E., and Jacques, J. (2015). The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*.
- Brinkmann, J., Ulmer, M. W., and Mattfeld, D. C. (2020). The multi-vehicle stochastic-dynamic inventory routing problem for bike sharing systems. *Business Research*, 13(1):69–92.
- Buck, D. and Buehler, R. (2012). Bike lanes and other determinants of capital bikeshare trips. In *91st Transportation research board annual meeting*, pages 703–706. Washington, DC.

- Bulhões, T., Subramanian, A., Erdoğan, G., and Laporte, G. (2018). The static bike relocation problem with multiple vehicles and visits. *European Journal of Operational Research*, 264(2):508–523.
- Caceres-Cruz, J., Arias, P., Guimarans, D., Riera, D., and Juan, A. A. (2014). Rich vehicle routing problem: Survey. *ACM Computing Surveys (CSUR)*, 47(2):1–28.
- Cahyono, M. R. A., Mariza, I., and Wirawan (2022). Electric bike monitoring and controlling system based on internet of things. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, 11(1):53–60.
- Callegati, F., Giallorenzo, S., Melis, A., and Prandini, M. (2018). Cloud-of-Things meets Mobility-as-a-Service: An insider threat perspective. *Computers and Security*.
- Casella, G. and Berger, R. (2024). *Statistical inference*. CRC press.
- Centro Nacional de Información Geográfica (CNIG) (2024). Web: Centro de descargas del CNIG. <https://centrodedescargas.cnig.es/CentroDescargas/>.
- Chemla, D., Meunier, F., and Wolfler Calvo, R. (2012). Bike sharing systems: Solving the static rebalancing problem. *Discrete Optimization*, 10(2):120–146.
- Chen, F., Turoń, K., Kłos, M., Czech, P., Pamuła, W., and Sierpiński, G. (2018). Scientific Journal of Silesian University of Technology . Series Transport FIFTH GENERATION OF BIKE-SHARING SYSTEMS – EXAMPLES OF POLAND AND CHINA. *Scientific Journal of Silesian University of Technology. Series Transport.*, 99:5–13.
- Chiariotti, F., Pielli, C., Zanella, A., and Zorzi, M. (2017). A dynamic approach to rebalancing bike-sharing systems. *Sensors (Switzerland)*, 18(2):1–22.
- Chiou, Y.-C. and Wu, K.-C. (2024). Bikesharing: The first-and last-mile service of public transportation? evidence from an origin–destination perspective. *Transportation Research Part A: Policy and Practice*, 187:104162.
- Contardo, C., Morency, C., and Rousseau, L.-M. (2012). *Balancing a dynamic public bike-sharing system*, volume 4. Cirrelt Montreal, Canada.
- Cruz, F., Subramanian, A., Bruck, B. P., and Iori, M. (2016). A heuristic algorithm for a single vehicle static bike sharing rebalancing problem. *Computers and Operations Research*, 79:19–33.
- Curtis, S. K. and Mont, O. (2020). Sharing economy business models for sustainability. *Journal of Cleaner Production*, 266:121519.
- Dantzig, G. B. and Ramser, J. H. (1959). The truck dispatching problem. *Management science*, 6(1):80–91.
- De Chardon, C. M., Caruso, G., and Thomas, I. (2016). Bike-share rebalancing strategies, patterns, and purpose. *Journal of transport geography*, 55:22–39.
- Dehaene, S. (2003). The neural basis of the weber–fechner law: a logarithmic mental number line. *Trends in cognitive sciences*, 7(4):145–147.
- del Río, S., Herrero, L., Fraile, R., and Penas, A. (2011). Spatial distribution of recent rainfall trends in spain (1961–2006). *International Journal of Climatology*, 31(5):656–667.

- Dell'Amico, M., Hadjicostantinou, E., Iori, M., and Novellani, S. (2014). The bike sharing rebalancing problem: Mathematical formulations and benchmark instances. *Omega (United Kingdom)*, 45:7–19.
- Dell'Amico, M., Iori, M., Novellani, S., and Stützle, T. (2016). A destroy and repair algorithm for the Bike sharing Rebalancing Problem. *Computers and Operations Research*, 71:149–162.
- Dell'Amico, M., Iori, M., Novellani, S., and Subramanian, A. (2018). The Bike sharing Rebalancing Problem with Stochastic Demands. *Transportation Research Part B: Methodological*, 118:362–380.
- Dell'Asin, G., Monzón, A., and Lopez-Lambas, M. E. (2015). Key quality factors at urban interchanges. In *Proceedings of the Institution of Civil Engineers-Transport*, volume 168, pages 326–335. Thomas Telford Ltd.
- DeMaio, P. (2009). Bike-sharing: History, Impacts, Models of Provision, and Future. *Journal of Public Transportation*, 12(4):41–56.
- DeMaio, P. (2007). Web: The Meddin Bike-sharing World Map. Accedido : Septiembre de 2024. <https://bikesharingworldmap.com>.
- Di Gaspero, L., Rendl, A., and Urli, T. (2013). Constraint-based approaches for balancing bike sharing systems. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8124 LNCS:758–773.
- Ding, D., Luo, M., Infante, M. F. P., Gunn, L., Salvo, D., Zapata-Diomed, B., Smith, B., Bellew, W., Bauman, A., Nau, T., et al. (2024). The co-benefits of active travel interventions beyond physical activity: a systematic review. *The Lancet Planetary Health*, 8(10):e790–e803.
- Du, M. and Cheng, L. (2018). Better understanding the characteristics and influential factors of different travel patterns in free-floating bike sharing: Evidence from nanjing, china. *Sustainability*, 10(4):1244.
- El-Assi, W., Salah Mahmoud, M., and Nurul Habib, K. (2017). Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in toronto. *Transportation*, 44:589–613.
- Elhenawy, M. and Rakha, H. (2017). A heuristic for rebalancing bike sharing systems based on a deferred acceptance algorithm. *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2017 - Proceedings*, pages 188–193.
- Empresa Municipal de Transportes de Madrid S. A. (EMT Madrid) (2024). Web: Principales cifras. [https://www.emtmadrid.es/Elementos-Cabecera/Enlaces-Pie-vertical/EMPRESA/Somos/Principales-cifras-\(1\)](https://www.emtmadrid.es/Elementos-Cabecera/Enlaces-Pie-vertical/EMPRESA/Somos/Principales-cifras-(1)).
- EMT Madrid (2024a). Bicimad. datos de los itinerarios de las bicicletas eléctricas. [https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-\(1\)](https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)) (accedido en 10 Jan 2023).
- EMT Madrid (2024b). Bicimad. inventario de las estaciones del servicio bicimad. <https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=e9b2a4059b4b7410VgnVCM2000000c205a0aRCRD&vgnnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnnextfmt=default> (accedido en 10 Jan 2023).

- EMT Madrid (2024c). Bicimad. modelo de datos.
<https://datos.madrid.es/FWPProjects/egob/Catalogo/Transporte/Bici/Ficheros/Servicios-y-estructuras-Bicimad-V1-1.pdf> (accedido en 10 Jan 2023).
- EMT Madrid (2024). Restapi emt estado de las estaciones de bicimad.
”<https://apidocs.emtmadrid.es/>”.
- Erdoan, G., Battarra, M., and Wolfler Calvo, R. (2015). An exact algorithm for the static rebalancing problem arising in bicycle sharing systems. *European Journal of Operational Research*, 245(3):667–679.
- Erdoğan, G., Battarra, M., and Calvo, R. W. (2015). An exact algorithm for the static rebalancing problem arising in bicycle sharing systems. *European Journal of Operational Research*, 245(3):667–679.
- Erdoğan, G., Laporte, G., and Calvo, R. W. (2014). The static bicycle relocation problem with demand intervals. *European Journal of Operational Research*, 238(2):451–457.
- Eren, E. and Uz, V. E. (2020). A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable cities and society*, 54:101882.
- Eurostat (2024). Web:Eurostat. Passenger mobility statistics.
https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Passenger_mobility_statistics.
- Fishman, E. (2015). Bikeshare: A Review of Recent Literature. *Transport Reviews*, 36(1):92–113.
- Fishman, E., Washington, S., and Haworth, Narelle, L. (2013). Bike Share. *Transport reviews*, 33(2):1–26.
- Fishman, E., Washington, S., and Haworth, N. (2012). Barriers and facilitators to public bicycle scheme use: A qualitative approach. *Transportation research part F: traffic psychology and behaviour*, 15(6):686–698.
- Fishman, E., Washington, S., and Haworth, N. (2014). Bike share’s impact on car use: Evidence from the united states, great britain, and australia. *Transportation research Part D: transport and environment*, 31:13–20.
- Forma, I. A., Raviv, T., and Tzur, M. (2015). A 3-step math heuristic for the static repositioning problem in bike-sharing systems. *Transportation research part B: methodological*, 71:230–247.
- Fricker, C. and Gast, N. (2016). Incentives and redistribution in homogeneous bike-sharing systems with stations of finite capacity. *Euro journal on transportation and logistics*, 5:261–291.
- Galatoulas, N.-F., Genikomsakis, K. N., and Ioakimidis, C. S. (2020). Spatio-temporal trends of e-bike sharing system deployment: A review in europe, north america and asia. *Sustainability*, 12(11):4611.
- Gao, L., Wang, J., and Wu, X. (2024). Promoting sustainable usage behavior in the sharing economy business model: A study based on bike-sharing. *Research in Transportation Business & Management*, 57:101241.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

- Godichon-Baggioni, A., Maugis-Rabusseau, C., and Rau, A. (2019). Clustering transformed compositional data using k-means, with applications in gene expression and bicycle sharing system data. *Journal of Applied Statistics*, 46(1):47–65.
- Golden, B., Wang, X., and Wasil, E. (2023). The evolution of the vehicle routing problem—a survey of vrp research and practice from 2005 to 2022. In *The Evolution of the Vehicle Routing Problem: A Survey of VRP Research and Practice from 2005 to 2022*, pages 1–64. Springer.
- Gong, W., Rui, J., and Li, T. (2024). Deciphering urban bike-sharing patterns: An in-depth analysis of natural environment and visual quality in new york’s citi bike system. *Journal of Transport Geography*, 115:103799.
- Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z., Azar, M. G., et al. (2020). koray kavukcuoglu, remi munos, and michal valko. bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.
- Guo, Y., Essa, M., Sayed, T., Haque, M. M., and Washington, S. (2019). A comparison between simulated and field-measured conflicts for safety assessment of signalized intersections in australia. *Transportation research part C: emerging technologies*, 101:96–110.
- Guo, Y., Yang, L., and Chen, Y. (2022). Bike share usage and the built environment: a review. *Frontiers in public health*, 10:848169.
- Guo, Y., Zhou, J., Wu, Y., and Li, Z. (2017). Identifying the factors affecting bike-sharing usage and degree of satisfaction in ningbo, china. *PloS one*, 12(9):e0185100.
- Guyader, H., Friman, M., and Olsson, L. E. (2021). Shared mobility: Evolving practices for sustainability. *Sustainability (Switzerland)*, 13(21):1–14.
- Hamari, J., Sjöklint, M., and Ukkonen, A. (2016). The sharing economy: Why people participate in collaborative consumption. *Journal of the association for information science and technology*.
- Han, L., Luong, B. T., and Ukkusuri, S. (2016). An algorithm for the one commodity pickup and delivery traveling salesman problem with restricted depot. *Networks and Spatial Economics*, 16(3):743–768.
- Hastie, T. (2009). The elements of statistical learning: data mining, inference, and prediction.
- Hinton, G. and Sejnowski, T. J. (1999). *Unsupervised learning: foundations of neural computation*. MIT press.
- Ho, A. (2014). Solving a static repositioning problem in bike-sharing systems using iterated tabu search. *Logistics and Transportation Review*, 69:180–198.
- Ho, S. C. and Szeto, W. Y. (2017). A hybrid large neighborhood search for the static multi-vehicle bike-repositioning problem. *Transportation Research Part B: Methodological*, 95:340–363.
- Hulot, P., Aloise, D., and Jena, S. D. (2018). Towards station-level demand prediction for effective rebalancing in bike-sharing systems. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 378–386.

- INE (Instituto Nacional de Estadística (2024). Web: Cifras oficiales de población resultantes de la revisión del Padrón municipal a 1 de enero. Accedido : Octubre de 2024. <https://www.ine.es/dynt3/inebase/es/index.htm?padre=517&capsel=525>.
- Inés, A., Domínguez, C., Heras, J., Mata, G., and Rubio, J. (2024). A topological approach for semi-supervised learning. *Journal of Computational Science*, 82:102403.
- Inhof, A. C., Cauchick-Miguel, P. A., Moro, S. R., and Zomer, T. T. d. S. (2024). Benchmarking bike-sharing systems: an analysis of the sustainable potential of use-oriented solutions. *Benchmarking: An International Journal*, 31(1):121–139.
- Instituto de Estadística (IE) de la Comunidad de Madrid (2024). Web: Estadísticas históricas. Climatología. Estación meteorológica de Retiro. <https://www.madrid.org/iestadis/fijas/estructu/general/otros/clehre.htm>.
- Jaguemont, J. and Bardé, F. (2023). A critical review of lithium-ion battery safety testing and standards. *Applied Thermal Engineering*, 231(June).
- Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.
- Jia, Y., Xu, Y., Yang, D., and Li, J. (2020). The Biobjective Bike-Sharing Rebalancing Problem with Balance Intervals: A Multistart Multiobjective Particle Swarm Optimization Algorithm. *Complexity*, 2020.
- Jiang, B., Yin, J., and Zhao, S. (2009). Characterizing the human mobility pattern in a large street network. *Phys. Rev. E*, 80:021136.
- Jittrapirom, P., Caiati, V., Feneri, A. M., Ebrahimigharehbaghi, S., Alonso-González, M. J., and Narayan, J. (2017). Mobility as a service: A critical review of definitions, assessments of schemes, and key challenges. *Urban Planning*.
- Juan and Willumsen, L. G. d. D. O. (2024). *Modelling transport*. John Wiley & Sons.
- Kamargianni, M., Li, W., Matyas, M., and Schäfer, A. (2016). A Critical Review of New Mobility Services for Urban Transport. *Transportation Research Procedia*, 14(0):3294–3303.
- Kloimüllner, C., Papazek, P., Hu, B., and Raidl, G. R. (2014). Balancing bicycle sharing systems: an approach for the dynamic case. In *Evolutionary Computation in Combinatorial Optimisation: 14th European Conference, EvoCOP 2014, Granada, Spain, April 23-25, 2014, Revised Selected Papers 14*, pages 73–84. Springer.
- Kloimüllner, C., Papazek, P., Hu, B., and Raidl, G. R. (2014). Balancing bicycle sharing systems: An approach for the dynamic case. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8600:73–84.
- Kölbl, R. and Helbing, D. (2003). Energy laws in human travel behaviour. *New Journal of Physics*, 5.
- Koller, D. (2009). Probabilistic graphical models: Principles and techniques.
- Kou, Z. and Cai, H. (2019). Understanding bike sharing travel patterns: An analysis of trip data from eight cities. *Physica A: Statistical Mechanics and its Applications*, 515:785–797.

- Krehbiel, J. N., Gabel, M. J., and Carrubba, C. J. (2017). The European court of justice. Transforming our world: the 2030 Agenda for Sustainable Development. *Routledge Handbook of Judicial Behavior*, 16301(October):467–490.
- Labadi, K., Benarbia, T., Hamaci, S., and Darcherif, A.-M. (2012). Petri nets models for analysis and control of public bicycle-sharing systems. *Petri Nets-Manufacturing and Computer Science*.
- Lazarus, J., Pourquier, J. C., Feng, F., Hammel, H., and Shaheen, S. (2020). Micromobility evolution and expansion: Understanding how docked and dockless bikesharing models complement and compete—a case study of san francisco. *Journal of Transport Geography*, 84:102620.
- Legros, B. (2019). Dynamic repositioning strategy in a bike-sharing system; how to prioritize and how to rebalance a bike station. *European Journal of Operational Research*, 272(2):740–753.
- Li, X., Zhang, Y., Du, M., and Yang, J. (2019a). Social factors influencing the choice of bicycle: Difference analysis among private bike, public bike sharing and free-floating bike sharing in kunming, china. *KSCE Journal of Civil Engineering*, 23:2339–2348.
- Li, Y., Szeto, W. Y., Long, J., and Shui, C. S. (2016). A multiple type bike repositioning problem. *Transportation Research Part B: Methodological*, 90:263–278.
- Li, Y., Zheng, Y., and Yang, Q. (2018). Dynamic Bike Reposition. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1724–1733.
- Li, Z., Zhang, J., Gan, J., Lu, P., Gao, Z., and Kong, W. (2019b). Large-scale trip planning for bike-sharing systems. *Pervasive and Mobile Computing*, 54:16–28.
- Liang, X., Zheng, X., Lv, W., Zhu, T., and Xu, K. (2012). The scaling of human mobility by taxis is exponential. *Physica A: Statistical Mechanics and its Applications*, 391(5):2135–2144.
- Lin, J.-H. and Chou, T.-C. (2012). A geo-aware and vrp-based public bicycle redistribution system. *International Journal of Vehicular Technology*, 2012(1):963427.
- Lin, J. R., Yang, T. H., and Chang, Y. C. (2013). A hub location inventory model for bicycle sharing system design: Formulation and solution. *Computers and Industrial Engineering*, 65(1):77–86.
- Liu, F., Lu, C., Gui, L., Zhang, Q., Tong, X., and Yuan, M. (2023). Heuristics for Vehicle Routing Problem: A Survey and Recent Advances. *arXiv:2303.04147*, pages 1–67.
- Liu, J., Sun, L., Chen, W., and Xiong, H. (2016). Rebalancing bike sharing systems: A multi-source data smart optimization. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Aug:1005–1014.
- Lu, M., Hsu, S.-C., Chen, P.-C., and Lee, W.-Y. (2018). Improving the sustainability of integrated transportation system with bike-sharing: A spatial agent-based approach. *Sustainable cities and society*, 41:44–51.
- Lu, Y., Benlic, U., and Wu, Q. (2020). An effective memetic algorithm for the generalized bike-sharing rebalancing problem. *Engineering Applications of Artificial Intelligence*, 95(August):103890.
- Maas, B. (2022). Literature Review of Mobility as a Service. *Sustainability (Switzerland)*.

- Mahajan, S. and Argota Sánchez-Vaquerizo, J. (2024). Global comparison of urban bike-sharing accessibility across 40 cities. *Scientific Reports*, 14(1):20493.
- Maleki, B., Mahdavi, I., Mahdavi-amiri, N., and Khorram, E. (2020). Balancing public bicycle sharing system using inventory critical levels in queuing network. *Computers & Industrial Engineering*, 141(January):106277.
- Marsland, S. (2011). *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC.
- Martínez-de Ibarreta, C., Cifuentes, J. A., Vallez, C. M., and Betancourt, A. (2024). Unveiling the effect of social media communication on urban mobility. *Transportation*.
- Mátrai, T. and Tóth, J. (2016). Comparative Assessment of Public Bike Sharing Systems. *Transportation Research Procedia*, 14:2344–2351.
- Mattson, J. and Godavarthy, R. (2017). Bike share in fargo, north dakota: Keys to success and factors affecting ridership. *Sustainable cities and society*, 34:174–182.
- McNeil, N., Broach, J., and Dill, J. (2018). Breaking barriers to bike share: Lessons on bike share equity. *Institute of Transportation Engineers. ITE Journal*, 88(2):31–35.
- Médard de Chardon, C., Caruso, G., and Thomas, I. (2016). Bike-share rebalancing strategies, patterns, and purpose. *Journal of Transport Geography*, 55:22–39.
- Metro de Madrid S.A (2023). Web: Metro de Madrid en cifras al cierre de 2023. <https://www.metromadrid.es/es/quienes-somos/metro-de-madrid-en-cifras>.
- Midgley, P. (2011). Bicycle-sharing schemes: Enhancing sustainable mobility in urban areas : United Nations Department of Economic and Social Affairs Commission on Sustainable Development Nine teenth Session New York, 2-13. *United Nations, Department of Economic and Social Affairs*.
- Mohiuddin, H., Fitch-Polse, D. T., and Handy, S. L. (2023). Does bike-share enhance transport equity? evidence from the sacramento, california region. *Journal of Transport Geography*, 109:103588.
- Mora, R., Miranda-Marquez, S., Truffello, R., and Sadarangani, K. P. (2024). Bikesharing and ordinary cyclists from chile: Comparing trips, attitudes, and health-behaviours. *Journal of Transport Geography*, 116:103826.
- Moshe E and Lerman, S. R. B.-A. (1985). *Discrete choice analysis: theory and application to travel demand*. MIT press.
- Mueller, N., Rojas-Rueda, D., Cole-Hunter, T., De Nazelle, A., Dons, E., Gerike, R., Götschi, T., Panis, L. I., Kahlmeier, S., and Nieuwenhuijsen, M. (2015). Health impact assessment of active transportation: a systematic review. *Preventive medicine*, 76:103–114.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
- Murphy, K. P. (2022). *Probabilistic machine learning: an introduction*. MIT press.
- Murphy, K. P. (2023). *Probabilistic machine learning: Advanced topics*. MIT press.

- Naciones Unidas (2015). Web: ODS Agenda 2030. Accedido : Julio de 2024.
<https://www.un.org/sustainabledevelopment/es/2015/09/la-asamblea-general-adopta-la-agenda-2030-para-el-desarrollo-sostenible/>.
- Nair, R. and Miller-Hooks, E. (2011). Fleet management for vehicle sharing operations. *Transportation Science*, 45(4):524–540.
- Nair, R., Miller-Hooks, E., Hampshire, R. C., and Bušić, A. (2013). Large-scale vehicle sharing systems: analysis of véliv'. *International Journal of Sustainable Transportation*, 7(1):85–106.
- Otero, I., Nieuwenhuijsen, M. J., and Rojas-Rueda, D. (2018). Health impacts of bike sharing systems in europe. *Environment international*, 115:387–394.
- Ou, Y., Bao, Z., Ng, S. T., and Song, W. (2023). Estimating the effect of air quality on bike-sharing usage in shanghai, china: An instrumental variable approach. *Travel Behaviour and Society*, 33:100626.
- O'Brien, O. (2024). Web: BIKE SHARE MAP. Global Map of Bikeshare by OOMap. Accedido : Septiembre de 2024. <https://bikesharemap.com>.
- O'Brien, O. and the Meddin Bike-sharing World Map team (2022). PDF: The Meddin Bike-sharing World Map Report 2022 edition. Accedido : Septiembre de 2024.
https://bikesharingworldmap.com/reports/bswm_mid2022report.pdf.
- Pérez-Cruz, F. (2008). Kullback-leibler divergence estimation of continuous distributions. *IEEE International Symposium on Information Theory - Proceedings*, pages 1666–1670.
- Pfrommer, J., Warrington, J., Schildbach, G., and Morari, M. (2014). Dynamic vehicle redistribution and online price incentives in shared mobility systems. *IEEE Transactions on Intelligent Transportation Systems*, 15(4):1567–1578.
- Poongodi, M., Malviya, M., Kumar, C., Hamdi, M., Vijayakumar, V., Nebhen, J., and Alyamani, H. (2022). New York City taxi trip duration prediction using MLP and XGBoost. *International Journal of System Assurance Engineering and Management*, 13(s1):16–27.
- Pucher, J., Dill, J., and Handy, S. (2010). Infrastructure, programs, and policies to increase bicycling: An international review. *Preventive medicine*, 50:S106–S125.
- Raidl, G. R., Hu, B., Rainer-Harbach, M., and Papazek, P. (2013). Balancing bicycle sharing systems: Improving a VNS by efficiently determining optimal loading operations. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7919 LNCS:130–143.
- Rainer-Harbach, M., Papazek, P., Hu, B., and Raidl, G. R. (2013a). Balancing bicycle sharing systems: A variable neighborhood search approach. In *Evolutionary Computation in Combinatorial Optimization: 13th European Conference, EvoCOP 2013, Vienna, Austria, April 3-5, 2013. Proceedings 13*, pages 121–132. Springer.
- Rainer-Harbach, M., Papazek, P., Hu, B., and Raidl, G. R. (2013b). Balancing bicycle sharing systems: A variable neighborhood search approach. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7832 LNCS:121–132.

- Raviv, T., Tzur, M., and Forma, I. A. (2013). Static repositioning in a bike-sharing system: models and solution approaches. *EURO Journal on Transportation and Logistics*, 2(3):187–229.
- Regue, R. and Recker, W. (2014a). Proactive vehicle routing with inferred demand to solve the bikesharing rebalancing problem. *Transportation Research Part E: Logistics and Transportation Review*, 72:192–209.
- Regue, R. and Recker, W. (2014b). Proactive vehicle routing with inferred demand to solve the bikesharing rebalancing problem. *Transportation Research Part E: Logistics and Transportation Review*, 72:192–209.
- Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. J., and Chong, S. (2011). On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking*, 19(3):630–643.
- Ricci, M. (2015). Bike sharing: A review of evidence on impacts and processes of implementation and operation. *Research in Transportation Business and Management*, 15:28–38.
- Rojas-Rueda, D., de Nazelle, A., Teixidó, O., and Nieuwenhuijsen, M. (2012). Replacing car trips by increasing bike and public transport in the greater barcelona metropolitan area: A health impact assessment study. *Environment International*, 49:100–109.
- Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.
- Samuel, A. L. (1959). Some Studies in Machine Learning. *IBM Journal of Research and Development*, 3(3):210–229.
- Santos, G. (2018). Sustainability and shared mobility models. *Sustainability (Switzerland)*, 10(9).
- Savelsbergh, M. W. and Sol, M. (1995). The general pickup and delivery problem. *Transportation science*, 29(1):17–29.
- Schuijbroek, J., Hampshire, R. C., and van Hoes, W. J. (2017). Inventory rebalancing and vehicle routing in bike sharing systems. *European Journal of Operational Research*, 257(3):992–1004.
- SciPy Community (2024). Scipy.stats: Statistical functions.
- Shaheen, S., Guzman, S., and Zhang, H. (2010). Bikesharing in Europe, the Americas, and Asia. *Transportation Research Record*, 2143(1):159–167.
- Shaheen, S., Martin, E., and Cohen, A. (2013). Public bikesharing and modal shift behavior: a comparative study of early bikesharing systems in north america. *International Journal of Transportation*.
- Shaheen, S. A., Zhang, H., Martin, E., and Guzman, S. (2011). China’s hangzhou public bicycle: understanding early adoption and behavioral response to bikesharing. *Transportation Research Record*, 2247(1):33–41.
- Shao, Z., Li, X., Guo, Y., and Zhang, L. (2020). Influence of service quality in sharing economy: Understanding customers’ continuance intention of bicycle sharing. *Electronic Commerce Research and Applications*, 40:100944.
- Shin, E. J. (2021). A comparative study of bike-sharing systems from a user’s perspective: An analysis of online reviews in three us regions between 2010 and 2018. *International journal of sustainable transportation*, 15(12):908–923.

- Shui, C. S. and Szeto, W. Y. (2017a). Dynamic green bike repositioning problem – A hybrid rolling horizon artificial bee colony algorithm approach. *Transportation Research Part D: Transport and Environment*, 60:119–136.
- Shui, C. S. and Szeto, W. Y. (2017b). Dynamic green bike repositioning problem – A hybrid rolling horizon artificial bee colony algorithm approach. *Transportation Research Part D: Transport and Environment*, 60:119–136.
- Si, H., gang Shi, J., Wu, G., Chen, J., and Zhao, X. (2019). Mapping the bike sharing research published from 2010 to 2018: A scientometric review. *Journal of Cleaner Production*, 213:415–427.
- Si, S., Chen, H., Liu, W., and Yan, Y. (2021). Disruptive innovation, business model and sharing economy: the bike-sharing cases in china. *Management Decision*, 59(11):2674–2692.
- Singleton, P. A. and Clifton, K. J. (2014). Exploring synergy in bicycle and transit use: Empirical evidence at two scales. *Transportation research record*, 2417(1):92–102.
- Sutton, R. S. (2018). Reinforcement learning: An introduction. *A Bradford Book*.
- Szeto, W. Y., Liu, Y., and Ho, S. C. (2016). Chemical reaction optimization for solving a static bike repositioning problem. *Transportation Research Part D: Transport and Environment*, 47:104–135.
- Tang, Q., Fu, Z., and Qiu, M. (2019). A Bilevel Programming Model and Algorithm for the Static Bike Repositioning Problem. *Journal of Advanced Transportation*, 2019.
- Tang, Q., Fu, Z., Zhang, D., Guo, H., and Li, M. (2020). Addressing the Bike Repositioning Problem in Bike Sharing System: A Two-Stage Stochastic Programming Model. *Scientific Programming*, 2020.
- Teusch, J., Gremmel, J. N., Koetsier, C., Johora, F. T., Sester, M., Woisetschlager, D. M., and Muller, J. P. (2023). A Systematic Literature Review on Machine Learning in Shared Mobility. *IEEE Open Journal of Intelligent Transportation Systems*, 4(October):870–899.
- The Comprehensive R Archive Network. (2024). Package rstan.
- Troped, P. J., Brenner, P. S., and Wilson, J. S. (2023). Associations between the built and social environment and bike share, physical activity, and overall cycling among adults from boston neighborhoods. *Journal of Transport & Health*, 31:101629.
- Vallez, C. M., Castro, M., and Contreras, D. (2021). Challenges and opportunities in dock-based bike-sharing rebalancing: A systematic review. *Sustainability (Switzerland)*, 13(4):1–26.
- Vallez, C. M., Contreras, D., and Castro, M. (2024). A probabilistic machine learning framework for modeling bicycle sharing demand: The bicimad case study. *Computers and Industrial Engineering*. Under review.
- Van Eck, N. and Waltman, L. (2010). Software survey: Vosviewer, a computer program for bibliometric mapping. *scientometrics*, 84(2):523–538.
- van Waes, A., Farla, J., Frenken, K., de Jong, J. P., and Raven, R. (2018a). Business model innovation and socio-technical transitions. a new prospective framework with an application to bike sharing. *Journal of Cleaner Production*, 195:1300–1312.

- van Waes, A., Farla, J., Frenken, K., de Jong, J. P., and Raven, R. (2018b). Business model innovation and socio-technical transitions. a new prospective framework with an application to bike sharing. *Journal of Cleaner Production*, 195:1300–1312.
- Varela, A F - Business Insider (2019). Web: Business Insider: 7 cifras sorprendentes sobre la economía de Madrid: PIB, riqueza per cápita, inversión extranjera o tejido empresarial. <https://www.businessinsider.es/7-cifras-sorprendentes-economia-madrid-415949>.
- Varga, L., Kovács, A., Tóth, G., Papp, I., and Nédá, Z. (2016). Further we travel the faster we go. *PLoS ONE*, 11(2):1–9.
- Vassallo, J. M., Tarriño, J., Gomez. Juan, and Soria-Lara, J. A. (2021). Impacto en la aceptabilidad y reparto modal de las medidas para mejorar la calidad del aire en Madrid central. *R-Evolucionando el transporte*, pages 1591–1627.
- Waltman, L., Van Eck, N. J., and Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of informetrics*, 4(4):629–635.
- Wang, B., Guo, Y., Chen, F., and Tang, F. (2024). The impact of the social-built environment on the inequity of bike-sharing use: A case study of divvy system in chicago. *Travel Behaviour and Society*, 37:100873.
- Wang, S. and Wu, R. (2019). The static rebalancing problem in bicycle-sharing systems with unusable bicycles. *2019 16th International Conference on Service Systems and Service Management, ICSSSM 2019*.
- Warrington, J. and Ruchti, D. (2019). Two-stage stochastic approximation for dynamic rebalancing of shared mobility systems. *Transportation Research Part C: Emerging Technologies*, 104(September 2018):110–134.
- Wikimedia Foundation, Inc. (2024). Stanisław ulam.
- Wood, J., Slingsby, A., and Dykes, J. (2011). Visualizing the dynamics of London’s bicycle-hire scheme. *Cartographica*, 46(4):239–251.
- Woodcock, J., Tainio, M., Cheshire, J., O’Brien, O., and Goodman, A. (2014). Health effects of the london bicycle sharing system: health impact modelling study. *Bmj*, 348.
- Wu, X., Lu, Y., Lin, Y., and Yang, Y. (2019). Measuring the destination accessibility of cycling transfer trips in metro station areas: A big data approach. *International journal of environmental research and public health*, 16(15):2641.
- Yan, X. Y., Han, X. P., Wang, B. H., and Zhou, T. (2013). Diversity of individual mobility patterns and emergence of aggregated scaling laws. *Scientific Reports*, 3:1–5.
- You, P. S. (2019). A two-phase heuristic approach to the bike repositioning problem. *Applied Mathematical Modelling*, 73:651–667.
- Zhang, D., Xu, W., Ji, B., Li, S., and Liu, Y. (2020). Computers and Operations Research An adaptive tabu search algorithm embedded with iterated local search and route elimination for the bike repositioning and recycling problem. *Computers and Operations Research*, 123:105035.

- Zhang, J., Zhang, L., Sun, F., and Wang, Z. (2018). An Overview on Thermal Safety Issues of Lithium-ion Batteries for Electric Vehicle Application. *IEEE Access*, 6:23848–23863.
- Zhang, L., Zhang, J., Duan, Z.-y., and Bryde, D. (2015). Sustainable bike-sharing systems: characteristics and commonalities across cases in urban china. *Journal of cleaner production*, 97:124–133.
- Zhang, R., Kan, H., Wang, Z., and Liu, Z. (2023). Relocation-related problems in vehicle sharing systems: A literature review. *Computers and Industrial Engineering*, 183(April):109504.
- Zhang, Y. (2017). *Bike-sharing usage : mining on the trip data of bike-sharing users*. University of Twente.
- Zhao, J., Deng, W., and Song, Y. (2014). Ridership and effectiveness of bikesharing: The effects of urban features and system characteristics on daily use and turnover rate of public bikes in China. *Transport Policy*, 35:253–264.
- Zhou, J., Guo, Y., Sun, J., Yu, E., and Wang, R. (2022). Review of bike-sharing system studies using bibliometrics method. *Journal of Traffic and Transportation Engineering (English Edition)*, 9(4):608–630.
- Zhou, X. (2015). Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in chicago. *PloS one*, 10(10):e0137922.