Universidades de Burgos, León y Valladolid

Máster universitario

Inteligencia de Negocio y Big Data en Entornos Seguros

TFM del Máster en Inteligencia de Negocio y Big Data en Entornos Seguros

Aplicación del método trip chaining sobre la red de transporte público de Madrid

Presentado por Carlos Cubo Izquierdo
en Universidad de Valladolid — 2 de septiembre de 2022
Tutores: Miguel Ángel Martínez Prieto
Jorge Silvestre Vilches

Resumen

En la actualidad, multitud de iniciativas de investigación están abordando los diferentes y variados temas relacionados con la implantación de un nuevo modelo de ciudades inteligentes, "*smart cities*", que involucran un mayor nexo entre la tecnología y la sociedad. Entre los muchos retos que presenta este concepto emergente se encuentra la gestión del tráfico urbano inteligente, que a su vez engloba el estudio de diversos campos de investigación relacionados.

En este trabajo trataremos el problema de la estimación de las paradas de bajada a través del método estadístico *trip chaining* para poder reconstruir los flujos de tránsito de pasajeros dentro de la red intermodal de transporte público de Madrid y construir la correspondiente matriz Origen-Destino (OD), la cual representa la frecuencia de los movimientos de pasajeros entre las distintas ubicaciones de subida y bajada que constituyen la red de transporte público de una ciudad.

Los resultados derivados de este proyecto demuestran la capacidad de *trip chaining* para dar una estimación de la parada de bajada en el 89,60 % de los segmentos de viaje procesados. Asimismo, presenta una tasa de acierto del 81,97 % entre los segmentos de viaje cuya parada de bajada puede ser validada por ser conocida, estableciendo como margen de error permitido 1 parada de ruta de distancia.

Finalmente, se presentan varias matrices OD de tránsito construidas a diferentes niveles de agrupación, que pueden ser de gran utilidad para poder realizar análisis más avanzados en el futuro.

Descriptores

Trip chaining, Estimación de la parada de bajada, Matriz Origen-Destino de tránsito, Transporte público intermodal, Smart cities.

Abstract

Nowadays, numerous research initiatives are addressing the different and varied issues related to the implementation of a new model of urban cities, "smart cities", which involve a stronger link between technology and society. Among the many challenges presented by this emerging concept is intelligent urban traffic management, which in turn encompasses the study of several related research fields.

In this thesis we will deal with the problem of alighting stop estamation through the statistical method trip chaining in order to reconstruct the passenger transit flows within the intermodal public transport network of Madrid and to construct the corresponding Origin-Destination (OD) matrix, which represents the frequency of passenger movements between the different boarding and alighting locations that constitute the public transport network of a city.

The results derived from this project demonstrate the ability of trip chaining to give an estimation of the alighting stop in 89,60 % of the processed trip segments. It also presents an accuracy of 81,97 % among the trip segments whose alighting stop can be validated by being known, setting an allowable margin of error of 1 distance route stop.

Finally, several transit OD matrices constructed at different grouping levels are presented, which may be useful for more advanced analyses in the future.

Keywords

Trip chaining, Alighting stop estimation, Transit Origin-Destination matrix, intermodal public transport, Smart cities.

Índice general

Índice de figuras

Índice de tablas

Memoria

Introducción

La constante evolución de las tecnologías a nuestro alcance junto con su inserción de pleno en la sociedad actual llevan consigo una serie de desafíos que deben ser abordados para promover la evolución de nuestra comunidad. En la actualidad, multitud de datos son recopilados con la realización de todo tipo de acciones por parte de personas, máquinas, sistemas, etc. El procesamiento y análisis de estos datos puede generar valor aplicable a muchos casos de uso distintos.

La emergente revolución de los datos fundamentada en el popular término *Big Data* y las diferentes propiedades relativas que lo rodean, las conocidas como V's del *Big Data*, requiere de acciones en muy diversos campos de estudio que permitan aprovechar todo lo que pueden ofrecernos los datos [14]. El tratamiento de estos datos masivos nos debe llevar a la consecución de nuevos desarrollos forjados a partir de ellos que den respuesta a las nuevas necesidades que surjan en la sociedad y que, sin duda, supondrán verdaderos retos a afrontar. Las V's del *Big Data* (véase la Figura 1.1) se corresponden con las características principales que describen los datos del nuevo ecosistema en el que nos encontramos actualmente. El número de V's consideradas es variable y ha evolucionado incrementalmente con el tiempo, hasta llegar a hablarse de la existencia de 42 distintas [15]. En este caso, resumiremos de forma concisa el significado de las 5 más extendidas.

- **Volumen**: describe la gran cantidad de información que se genera.
- **Velocidad**: describe la alta rapidez con la que se generan y consumen los datos.
- **Variedad**: describe las diversas formas que pueden adoptar los datos.
- **Valor**: describe la aptitud de los datos para aportar valor en un contexto de negocio.
- **Veracidad**: describe el grado de fiabilidad de los datos.

Entre los retos necesarios y de mayor impacto sobre la vida diaria de las personas destaca la implantación de un nuevo modelo de ciudades Inteligentes, también denominadas

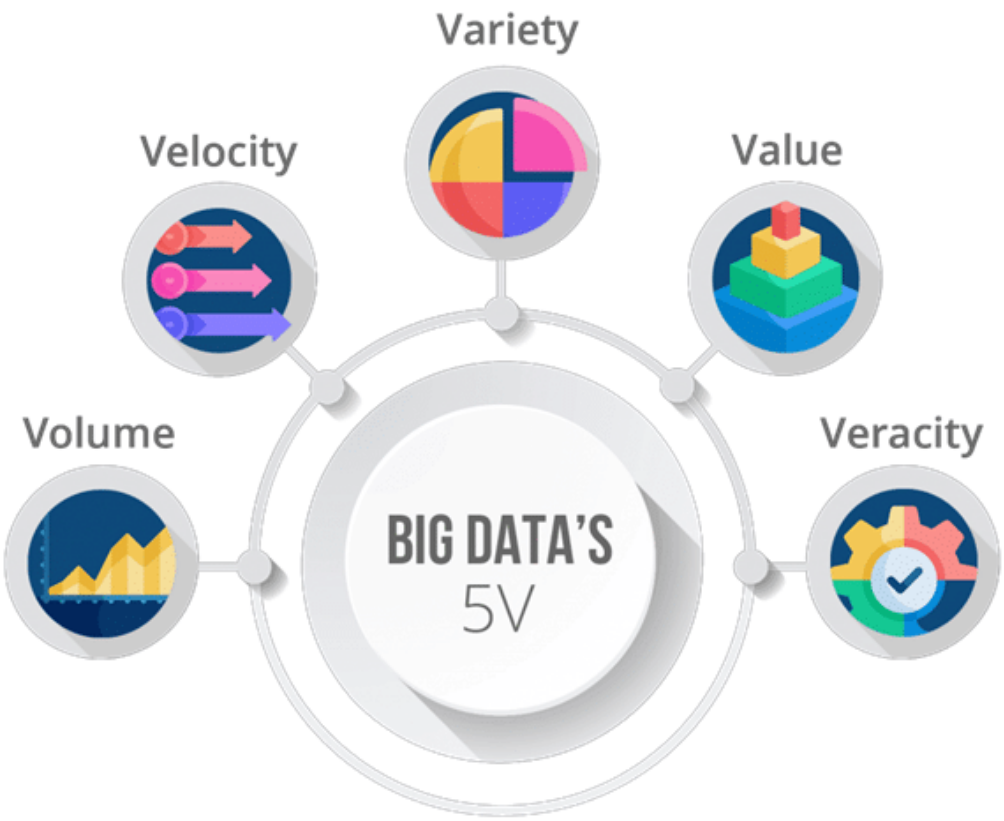

Figura 1.1: Las principales 5 V's del Big Data [1].

"*smart cities*". Este concepto de ciudad inteligente refiere a un nuevo tipo de desarrollo urbano que, basado en las ideas de sostenibilidad y eficiencia, propone cambiar el funcionamiento general de las ciudades del futuro [16]. Algunas de las múltiples áreas de actuación dentro de estas ciudades inteligentes son [17, 18, 2]: transporte y movilidad, eficiencia energética, telecomunicaciones, servicios sociales y de salud, educación, gestión de residuos, gestión de redes de abastecimiento, seguridad ciudadana o recaudación de impuestos, tal cual se ilustra en la Figura 1.2. En nuestro caso, nos centraremos en el área de aplicación relativo al *transporte y la movilidad* referido como **transporte urbano inteligente**. En concreto, trataremos un problema relacionado con la gestión del tráfico urbano como es la reconstrucción de los flujos de viaje de los pasajeros dentro de una red de transporte público.

En la mayoría de los casos, la problemática relativa a esta reconstrucción de viajes se debe al desconocimiento de la ubicación de bajada en la que los pasajeros salen del medio de transporte público utilizado, siendo conocida habitualmente la ubicación de su subida, al entrar en la red de transporte público. A la hora de abordar la estimación de la parada

Figura 1.2: Áreas de Actuación en *Smart Cities* [2].

de bajada (*alighting stop estimation*) de los pasajeros en sus viajes en el transporte público existen diversas aproximaciones válidas, entre las que destaca el método *trip chaining* [19].

En este trabajo se propone la aplicación de una adaptación de este método estadístico basado en reglas para la determinación de las paradas de bajada de los pasajeros en sus viajes realizados en la red de transporte público de Madrid. Otra cuestión relevante a considerar es la detección de los transbordos entre vehículos de transporte que sirven a los pasajeros para moverse de un modo de transporte de la red a otro o, simplemente, para cambiar de ruta en un mismo modo. Es importante remarcar que la parada de transbordo de un segmento de viaje no debe ser considerada como la parada de bajada del viaje, sino como una parada intermedia y necesaria, que toma el pasajero en su camino hacia el destino deseado. A este respecto, se dispondrá de un conjunto de reglas empleadas como mecanismo para determinar si se ha producido una actividad final, lo cual implicará la terminación del viaje, o un transbordo entre la parada de bajada de un segmento de viaje y la de subida del subsiguiente segmento de viaje del pasajero. También, se propone la construcción de una matriz Origen-Destino (*OD matrix*) de tránsito que represente la frecuencia de ocurrencia de cada posible trayecto de viaje constituido por sus ubicaciones de origen (subida) y destino (bajada) dentro de la red de transporte público de una ciudad.

1.1. Motivación

Las aplicaciones en el dominio de las ciudades inteligentes son un campo emergente y de mucho futuro, al abarcar un amplio espectro de áreas de innovación que influyen en nuestro día a día. En concreto, los temas relacionados con la gestión del tráfico inteligente se sitúan entre los más importantes y, en consecuencia, deben ser abordados. El acuciante movimiento por el cambio climático también ha construido a un mayor interés por parte de la comunidad en emprender la innovación en estas áreas de investigación, de cara a la mejora de la eficiencia y sostenibilidad de los procesos relativos al tráfico [20].

En el contexto que nos ocupa, la reconstrucción de los flujos de viaje de pasajeros en la red de transporte público de una ciudad (como medio para la determinación de su matriz OD de tránsito) se presenta como una tarea inicial de gran importancia a la hora de poder disponer de un mayor conocimiento acerca de la dinámica de viajes que se producen en la red de transporte considerado. La Figura 1.3 presenta un ejemplo de matriz OD de tránsito. En ella se representan las diferentes zonas de paradas de la red de transporte de Oporto, replicadas como posibles origen (eje Y) y destino (eje X) de los viajes de los pasajeros. Cada una de las intersecciones de la matriz indica la frecuencia de viajes de cada par origen-destino de zonas, cuyo valor, representado a través de una escala de color gradual, representa el volumen de viajes ocurrido entre ambas zonas de la red. Así, los pares origen-destino más frecuentes se expresan con un color más intenso y los menos frecuentes con un color menos intenso.

La obtención de esta valiosa información habilita la realización posterior de análisis avanzados, en temas de optimización y planificación del tráfico en estas ciudades inteligentes. Por tanto, la estimación efectiva de las matrices OD de tránsito (que describen el uso de las redes de transporte de pasajeros) motiva el desarrollo de casos de uso que, tomando estas matrices como entrada, resultan en grandes progresos en multitud de campos. Algunos de estos casos de uso de interés son los siguientes:

- Planificación de horarios y rutas de transporte según la demanda.
- Asignación y distribución de recursos de la red de transporte.
- Modelado de redes de transporte a gran escala.
- Gestión operativa del tráfico en tiempo real.
- Evaluación de la congestión de paradas y estaciones.
- Adecuación de las infraestructuras de transporte a la demanda requerida.
- Priorización de financiación en proyectos relativos al tráfico.
- Identificación de inconsistencias entre la oferta y la demanda de servicios de transporte.

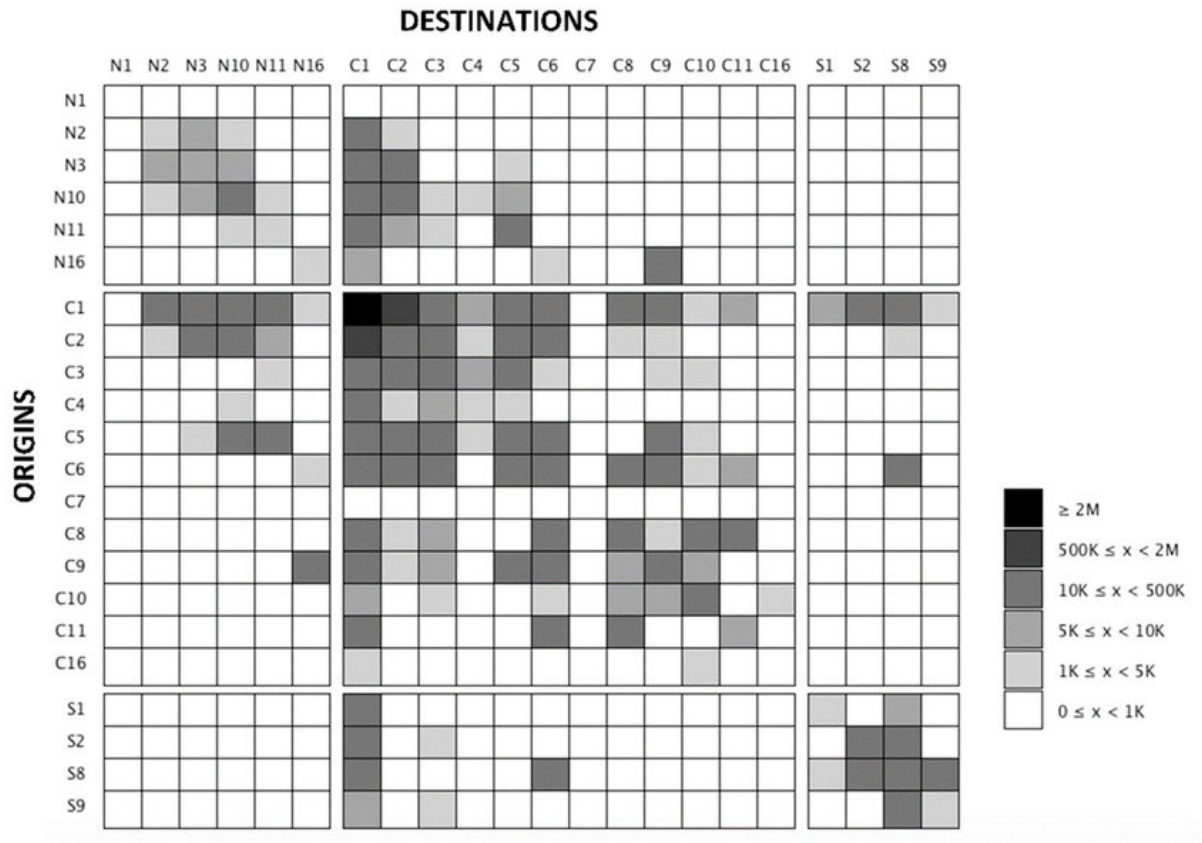

Figura 1.3: Ejemplo de Matriz OD de tránsito [3].

En la mayoría de los casos, las cuestiones a tratar en el contexto del transporte urbano inteligente son dependientes del entorno de negocio en el que se trabaja. Dentro de este proyecto, la obtención de las mencionadas matrices OD de tránsito involucra una serie de problemas que deben ser abordados previamente para su efectiva construcción, como la estimación de las paradas de subida y de bajada de los pasajeros o la detección de transbordos entre vehículos de transporte de la red. En este ámbito, el planteamiento para abordar la resolución de estas cuestiones se debe ajustar a las características y particularidades del contexto de la red de transporte pública concreta sobre la que se desarrolla el estudio. En nuestro proyecto, trabajaremos en el contexto de la red de transporte público de Madrid.

La red de transporte público de Madrid es una red intermodal de transporte de pasajeros constituida por 4 tipos principales de modos de transporte (metro, autobuses urbanos de la ciudad de Madrid, autobuses interurbanos, cercanías) que operan conjuntamente tanto en la Comunidad como en la ciudad de Madrid. El registro de las operaciones de viaje en esta red es realizado a través de terminales de un sistema automático de tarificación (AFC) instalados a lo largo de la misma. La mayor parte de este sistema AFC es de tipo

Figura 1.4: Mapa mundial con las principales ubicaciones en las que se han abordado problemas asociados a matrices OD [4].

entry-only, lo cual implica el registro de únicamente las transacciones de subida de los pasajeros, careciéndose de la información de bajada del transporte público.

Para la actual configuración de la red intermodal de transporte público de Madrid no se ha encontrado ningún estudio previo relevante que lleve a cabo el proceso de construcción de las matrices OD o que trate la reconstrucción de los flujos de viaje de pasajeros, representados por medio de su tarjeta personal de transporte, a partir de las transacciones de subida al transporte público que se recogen en los terminales del sistema de registro AFC (*Automatic Fare Collection*) instalados a lo largo de la red de transporte. La figura 1.4 muestra un mapa con las ubicaciones alrededor del mundo donde se han concretado las principales iniciativas de investigación que abordan este problema utilizando datos de transacciones de viaje de las redes de transporte público de estos lugares. Así, el presente trabajo surge para abordar esta necesidad en el caso del transporte público de Madrid.

De acuerdo con las características de la red de transporte público de Madrid, el principal problema de transporte urbano inteligente en el que nos centraremos será la estimación de las paradas de bajada de los pasajeros en sus viajes dentro de la red. En este trabajo, se propone la aplicación del método estadístico *trip chaining* para la estimación de dichas paradas de bajada junto con una serie de reglas básicas para la detección de transbordos con el objetivo de reconstruir los flujos de viaje de los pasajeros a través del encadenamiento de los segmentos de viaje (trip legs), representados por registros de transacciones en el transporte público, que constituyen los viajes completos origen-destino recorridos por los pasajeros para llegar desde su punto de partida inicial hasta su destino deseado.

Así, la principal motivación de este trabajo puede ser sintetizada en la siguiente pregunta de investigación: «**¿es viable la aplicación del método *trip chaining* sobre la red intermodal de transporte público de Madrid para construir las matrices OD de tránsito?**»

1.2. Objetivos y Restricciones del Proyecto

En este apartado se ofrece una descripción de los objetivos finales pretendidos con la realización de este trabajo relativo al transporte urbano inteligente. Así mismo, se listan una serie de restricciones que influirán en el alcance y la forma de realización de las tareas que llevarán al cumplimiento de los objetivos planteados.

En primer lugar, en la Tabla 1.1 se identifican y describen los principales objetivos que se plantean con la realización de ese proyecto.

ID Objetivo	**Nombre Objetivo**	**Descripción Objetivo**
OBJ-01	Revisión de estudios anteriores que tratan el problema de la construcción de matrices OD de tránsito en redes de transporte público	Consiste en el estudio del arte de artículos publicados anteriormente acerca de problemas enmarcados dentro la construcción de matrices OD de tránsito que reflejen la dinámica de flujos de viaje de pasajeros dentro de una red de transporte público.
OBJ-02	Análisis y transformación de los datos de transacciones y generales de la red de transporte público de Madrid	Consiste en la exploración previa y análisis posterior de los datos de registro de las transacciones producidas en las ubicaciones de la red de transporte público de Madrid, así como de los datos generales proporcionados por el Consorcio Regional de Transportes de Madrid (CRTM) para describir su topología y servicios ofrecidos. Además, se engloba el proceso ETL requerido para transformar los datos disponibles desde su forma en crudo hasta la adaptación para su uso efectivo.
OBJ-03	Aplicación del método *trip chaining* para la estimación de las paradas de bajada de los pasajeros	Consiste en el proceso de desarrollo necesario para adaptar el método *trip chaining* de estimación de las paradas de bajada de los pasajeros al caso particular de la red de transporte público de Madrid.

Tabla 1.1: Objetivos del Proyecto

Por otro lado, a continuación en la Tabla 1.2 se listan una serie de restricciones importantes a tener en cuenta de cara al desarrollo del proyecto.

ID Restricción	Nombre Restricción	Descripción Restricción
RES-01	Ausencia de datos de horarios concordantes con las transacciones	Implica que no se considerarán los datos horarios de la agenda de viajes, correspondientes a las horas de llegada y salida de los vehículos de transporte a cada parada de la red, a la hora de aplicar el método *trip chaining* para la estimación de las paradas de bajada de los pasajeros en el transporte público. Esto se debe a la no disponibilidad de los datos de horario relativos a las fechas en las que están comprendidas las transacciones de viaje de las que disponemos.
RES-02	Incapacidad de validación endógena	Implica que no se podrá validar de forma exacta cada una de las paradas de bajada estimadas mediante el método *trip chaining* al no disponer de esta información en el propio conjunto de transacciones a nuestra disposición. Hay que tener en cuenta que el sistema de registro de las transacciones de viaje instalado en la mayor parte de las ubicaciones de la red de transporte público es de tipo *entry-only* (ver definición en la Sección 2.1.2). No obstante, se hará una evaluación parcial del método *trip chaining* utilizando aquellos viajes de los cuales sí conocemos la bajada.

Tabla 1.2: Restricciones del Proyecto

1.3. Estructura del Documento

... Esta sección describe la organización del presente documento en torno a capítulos y resume el contenido de cada uno de ellos. Este documento se encuentra estructurado de la siguiente forma:

1. **Introducción**: en este primer capítulo se ofrece una breve introducción al entorno de negocio general en el que se encuadra el proyecto y se motiva el interés en su realización. Además, se aporta una descripción de los objetivos y restricciones principales del proyecto.

2. **Background**: en este capítulo se aporta una visión más completa sobre el contexto del problema concreto a tratar en el proyecto así como sobre las posibilidades para abordarlo que se consideran en la literatura. También, se describe de forma general la propia red de transporte público de Madrid en la que se centra el trabajo. Finalmente, el capítulo concluye con una revisión de las propuestas anteriores más destacadas que tratan el problema objeto de estudio.

3. **Planificación y Presupuesto**: en este capítulo se resume la planificación seguida para la realización del proyecto junto con una estimación de su presupuesto económico. También, incluye la descripción de las metodologías de trabajo y desarrollo adoptadas.

4. **Análisis**: este capítulo del proyecto comprende la descripción la etapade análisis del ciclo de vida del proyecto donde se realizan el conjunto de actividades destinadas a consolidar una abstracción del proyecto. Principalmente, se incluyen las tareas preliminares de exploración, estudio y modelado de los datos necesarias para conformar una primera aproximación a la solución del proyecto.

5. **Proceso ETL**: en este capítulo se lleva a cabo la implementación del diseño de transformación de datos descrito en el análisis del proyecto. Aporta una descripción detallada del proceso seguido para convertir los datos iniciales disponibles (*raw data*) en datos refinados para cumplir con los propósitos del proyecto (*smart data*).

6. **Implementación del método *trip chaining***: este capítulo aporta una explicación extendida de la adaptación tomada para aplicar el método *trip chaining* sobre la red de transporte público de Madrid. Comprende el proceso completo implementado para obtener las etapas de viaje y los viajes reconstruidos de los pasajeros de la red de transporte a partir de los datos de transacciones asociados a sus tarjetas de transporte.

7. **Evaluación**: en este capítulo se lleva a cabo la evaluación de los resultados obtenidos en el proyecto y la presentación de los más detacados en forma de gráficos y mapas. También, se explica el proceso de validación parcial seguido para estudiar el rendimiento del método *trip chaining* en la estimación de las paradas de bajada.

8. **Conclusiones y Trabajo Futuro**: este último capítulo expone una serie de conclusiones finales tanto a nivel de proyecto como de la aplicación del método *trip chaining* sobre la red de transporte de Madrid. Además, define varias líneas de trabajo futuro con las que poder continuar la investigación y desarrollo iniciados en este proyecto.

Background

En este capítulo se dará un contexto más extenso a los conceptos de fondo que se ven involucrados al introducirnos en este entorno de negocio de este proyecto. Se comienza presentando el concepto de transporte urbano inteligente, sus áreas de aplicación y las principales cuestiones a resolver acerca del problema tratado en este trabajo. Seguidamente, se describirá la estructura de la red de transporte público de Madrid, nuestro entorno de trabajo particular, y se explorarán diversos aspectos relativos a la misma. Finalmente, se presentará un estado del arte donde se revisan y comparan algunas de las propuestas anteriores más destacadas que tratan la resolución del problema objeto de estudio.

2.1. Transporte Urbano Inteligente

El transporte urbano inteligente es un campo de estudio ubicado dentro del área de investigación de las ciudades inteligentes, o *smart cities*, que consiste en la integración de la tecnología dentro de la gestión del tránsito en las ciudades, con el objetivo de optimizar las operaciones de planificación y control del tráfico y promover una movilidad más eficiente, segura y sostenible adaptada a las necesidades de la población [21].

Para tratar el estudio y desarrollo de nuevas formas y modelos de gestión del tráfico inteligente para su adaptación a esta nueva dimensión de las ciudades, en la que la tecnología es el componente destacado, se ve necesario el uso de datos de diverso tipo que nos permitan modelar mejor la compleja realidad. La creciente disponibilidad de cantidades masivas de nuevos datos procedentes del Internet de las Cosas (IoT), de redes de comunicaciones 5G, de sistemas de geoposicionamiento en tiempo real (AVL, GPS) o de sistemas automáticos de recogida de información (ADC), supone una gran oportunidad para aprovechar el potencial valor extraíble de los datos y obtener nuevo conocimiento, que antes no estaba a nuestra disposición.

2.1.1. Matriz OD

A continuación, se dará un mayor contexto al concepto de matriz OD de tránsito, al ser uno de los resultados de mayor interés en este proceso que posibilita la realización de análisis dentro del campo del transporte urbano inteligente.

La matriz Origen-Destino de tránsito es una representación de datos en forma matricial que permite modelar la dinámica de viajes dentro de una red de transporte de pasajeros. Su estructura axial en filas y columnas posibilita representar las ubicaciones de subida y bajada de pasajeros dentro de la red de transporte cuyas intersecciones indican la frecuencia de ocurrencia de cada combinación origen-destino posible. De esta forma, mientras sus filas y columnas se corresponden con las ubicaciones de origen y destino de viajes, su intersección representa la actividad de tránsito que se produce entre ellas, la cual puede ser estudiada a diferentes niveles de agregación, tanto espacial (e.g. *análisis por zonas*) como temporal (e.g. *análisis en horas punta*).

Una consideración importante a tener en cuenta a la hora de construir estas matrices OD es la detección de transbordos, los cuales permitirán el encadenamiento de segmentos de viaje realizados por un mismo pasajero con el propósito de llegar a una ubicación concreta, es decir, evitando estimar la parada de transbordo como el final de su viaje. De esta forma, las paradas de transbordo no serán consideradas como destino de viaje, evitando su representación en la matriz OD. Así, la consideración de los transbordos se presenta como un aspecto crucial para poder distinguir las paradas comunes de transbordo de las paradas comunes de destino y, por tanto, no crear matrices OD que den lugar a equívocos.

Como ya se comentó, las matrices OD de tránsito suponen una fuente de información muy valiosa de cara a realizar análisis posteriores más avanzados sobre otros retos variados del área del transporte urbano inteligente con el objetivo de estudiar diversas cuestiones que se deseen evaluar sobre la red de transporte considerada.

2.1.2. Datos

Centrándonos en la resolución del problema de la construcción de matrices OD de tránsito para redes de transporte público, los tipos y fuentes de datos utilizados en la literatura son variados, siendo su uso dependiente de la información concreta que se necesite obtener en cada caso. Algunos de los datos más comunes empleados en propuestas anteriores para resolver diferentes cuestiones relativas al problema se presentan a continuación.

Transacciones de viaje. Las transacciones de viaje representan la principal fuente de datos a utilizar para la reconstrucción de los flujos de viaje de los usuarios del transporte público. El registro de esta información se produce en los sistemas automáticos de tarifica-

ción (AFC) instalados en las distintas ubicaciones de la red de transporte al paso de los viajeros.

Dependiendo del modo de operación de la red de transporte público considerada, estas transacciones de viaje se registrarán en diferentes momentos de la marcha. Este hecho afectará a la localización de los terminales AFC dispuestos en las ubicaciones de subida/bajada de pasajeros en la red de transporte.

Principalmente, se consideran 3 tipos de sistemas AFC de registro de transacciones:

- ***Entry-only***: Sistema en el que los pasajeros pasan su tarjeta personal de transporte por los terminales habilitados solo en sus subidas al transporte público. Por lo tanto, las transacciones de viaje que se registran solo recogen la información de subida de los pasajeros, no teniendo información alguna sobre su bajada. Éste es el tipo de sistema AFC más habitual que se presenta en multitud de redes de transporte público de ciudades alrededor del mundo, como la de Madrid.
- ***Entry-exit***: Sistema en el que los pasajeros pasan su tarjeta personal de transporte por los terminales habilitados cada vez que entran o salen del transporte público. Por lo tanto, las transacciones de viaje que se obtienen serán correspondientes tanto a la subida como a la bajada de los pasajeros. Este tipo de sistema AFC de registro de transacciones abre un mayor abanico de posibilidades a la hora de utilizar métodos de estimación más avanzados, cuyo resultado puede ser validado sobre los propios datos de transacciones disponibles. A pesar de no ser este tipo de sistema el más habitual, existen numerosos estudios destacados desarrollados en estas circunstancias, sobre todo tomando como referencia la red de transporte público de South East Queensland, Australia [22, 23, 6]. Generalmente, los estudios de este tipo obvian la información de subida o bajada de los pasajeros con miras a su uso en una validación endógena, enfocada en comprobar la precisión en la estimación de esa información por parte de los métodos que se utilicen.
- ***Exit-only***: Sistema en el que los pasajeros pasan su tarjeta personal de transporte por los terminales habilitados solo en sus bajadas del transporte público. Por lo tanto, las transacciones de viaje que son registradas solo aportarán información sobre la salida de los pasajeros, no aportando información alguna sobre su subida. Éste es el tipo de sistema AFC menos habitual y es utilizado, especialmente, en redes de transporte donde los pasajeros parten desde ubicaciones en el centro de la ciudad para llegar a diferentes ubicaciones en la periferia cuya tarifa dependerá de la distancia recorrida desde el punto de partida.

El establecimiento en la red de transporte de un sistema AFC u otro implicará la disponibilidad de diversa información inicial, lo cual provocará la falta de cierta información en cada caso, constituyendo un tipo de problema de estimación concreto y, por lo tanto, el uso de técnicas y/o suposiciones diferentes para estimar la información restante de los viajes necesaria.

Las transacciones de viaje recopiladas por ambos tipos de sistema AFC suelen incluir datos similares para describir el suceso de la etapa del viaje concreta, tales como: ID Transacción, ID Tarjeta Transporte, ID Ubicación, Fecha y Hora, Modo de Transporte, etc. Igualmente, los atributos de información aportados por cada tipo de transacción también pueden variar dependiendo del modo de transporte en cuestión (autobús, metro, tren...) o tipo de ubicación dentro de la red de transporte donde se registre la transacción (estación, parada, vehículo...). Así, otros datos relevantes como la ruta y el sentido de ruta podrían estar presentes en las transacciones de viaje.

Datos GTFS. GTFS (*General Transit Feed Specification*) es un estándar abierto de Google para la especificación de información general de tránsito de transporte que permite a las agencias encargadas de la gestión de una red de transporte público compartir información diversa sobre los servicios que ofrece [24]. Los datos GTFS son conjuntos de datos adheridos a esta especificación, a través de los cuales las agencias proporcionan datos estáticos y/o dinámicos variados, tales como las rutas de transporte, los horarios de viaje establecidos, las tarifas asociadas, la designación de las paradas junto con su localización geográfica, además de información adicional de utilidad para modelar el funcionamiento de una red de servicios de transporte. Estos conjuntos de datos constituyen una fuente de interés a la hora de disponer de diversa información estructurada, que puede ser integrada con los datos de transacciones para ampliar el conocimiento de la realidad actual de viajes y poder abordar distintos problemas relativos al transporte urbano inteligente con más garantías.

El objetivo final de GTFS es establecer un formato común con el que poder hacer pública y usable dicha información para empresas externas, investigadores y desarrolladores, los cuales puedan crear nuevas aplicaciones para consumidores u otros desarrollos novedosos nutriéndose de esta información. El estándar GTFS está dividido en 2 componentes diferenciados: *GTFS Schedule*, con información estática de la red como horarios, tarifas o ubicaciones, y *GTFS Realtime*, con información dinámica en tiempo real como tiempos de predicción de llegada, posicionamiento de vehículos o avisos de servicio.

La información más relevante de cara al problema de la reconstrucción de flujos de viaje pasados es la proporcionada en el componente *GTFS Schedule*, enfocado en los datos estáticos de la red de transporte [25]. Las agencias de transporte que deseen adherirse a este estándar y compartir públicamente su información deberán crear conjuntos de datos siguiendo apropiadamente la especificación GTFS habilitada. Así, cada uno de estos conjuntos de datos GTFS, denominados *feeds*, deberá ser un fichero comprimido (ZIP) compuesto por una serie de ficheros cuyos registros se almacenan línea a línea delimitando los valores de sus campos por comas. Cada uno de los ficheros de texto que constituyen el feed GTFS, que pueden ser obligatorios u opcionales, describen un aspecto particular de la información de tránsito y cuentan con una determinada estructura de campos que deben satisfacer. La descripción de cada uno de estos ficheros se presenta en la Tabla 2.3. Igualmente, a modo de ejemplo, en las Tablas 2.4 y 2.5 se muestra un

resumen de los campos que componen los ficheros *stops.txt* y *routes.txt* dentro del *feed* GTFS correspondiente al Metro de Madrid.

Nombre Fichero	**Obligatoriedad**	**Descripción Fichero**
agency.txt	Obligatorio	Definición de las empresas de transporte.
stops.txt	Obligatorio	Definición de las paradas de subida y bajada de pasajeros.
routes.txt	Obligatorio	Definición de las rutas de transporte público.
trips.txt	Obligatorio	Especificación de los viajes asociados a cada ruta de transporte.
stop_times.txt	Obligatorio	Especificación de las horas de llegada y salida de los vehículos en las paradas de cada viaje.
calendar.txt	Condicionalmente obligatorio	Definición de las fechas de servicio asociadas a los viajes.
calendar_dates.txt	Condicionalmente obligatorio	Indicación de las excepciones a las fechas de servicio definidas en calendar.txt.
fare_attributes.txt	Opcional	Definición de la información sobre las tarifas de las rutas.
fare_rules.txt	Opcional	Definición de las reglas aplicadas sobre las tarifas de itinerarios de rutas.
shapes.txt	Opcional	Definición de las reglas para asignar las rutas de viaje de los vehículos, es decir, los alineamientos de rutas.
frecuencies.txt	Opcional	Indicación del tiempo entre viajes de un servicio.
transfers.txt	Opcional	Definición de reglas para el establecimiento de conexiones en los puntos de transbordo entre rutas.
pathways.txt	Opcional	Definición de los recorridos de conexión entre ubicaciones dentro de las estaciones.
levels.txt	Opcional	Indicación de los niveles de altura dentro de las estaciones.
feed_info.txt	Condicionalmente obligatorio	Inclusión de metadatos sobre el conjunto de datos.
translations.txt	Opcional	Traducción de los valores de campos en otros ficheros del conjunto de datos.
attributions.txt	Opcional	Especificación de las atribuciones aplicadas al conjunto de datos.

Tabla 2.3: Ficheros de la especificación GTFS Schedule

Campo	**Tipo**	**Obligatorio**	**Descripción**	**Ejemplo (Madrid)**
stop_id	ID	✓	Identificador de la ubicación o parada.	par_4_12
stop_code	Texto	✗	Texto breve o número asociado a la ubicación.	12
stop_name	Texto	Condicional	Nombre de la ubicación.	SOL
stop_desc	Texto	✗	Descripción de la ubicación.	Plaza de la Puerta del Sol 6
stop_lat	Latitud	Condicional	Latitud geográfica de la ubicación.	40.41688
stop_lon	Longitud	Condicional	Latitud geográfica de la ubicación.	-3.70326
zone_id	ID	Condicional	Identificador de la zona tarifaria de la ubicación.	A
stop_url	URL	✗	URL de una página asociada a la ubicación.	http:// www.crtm.es
location_type	Enum	✗	Tipo de la ubicación.	0
parent_station	ID en stops.stop_id	Condicional	Ubicación que contiene la actual según una jerarquía entre ubicaciones.	est_90_58
stop_timezone	Zona horaria	✗	Zona horario de la ubicación.	-
wheelchair_boarding	Enum	✗	Posibilidad de acceso en silla de ruedas a la ubicación.	1
level_id	ID en levels.level_id	✗	Nivel de la ubicación.	-
platform_code	Enum	✗	Identificador de la plataforma en la que se encuentra la ubicación.	-

Tabla 2.4: Campos del fichero *stops.txt* de la especificación GTFS Schedule

Campo	**Tipo**	**Obligatorio**	**Descripción**	**Ejemplo (Madrid)**
route_id	ID	✓	Identificador de la ruta.	4___1____
agency_id	ID en agency. agency_id	Condicional	Identificador de la empresa que opera la ruta.	CRTM
route_short_name	Texto	Condicional	Nombre corto de la ruta.	1
route_long_name	Texto	Condicional	Nombre completo de la ruta.	Pinar de Chamartín-Valdecarros
route_desc	Latitud	✗	Descripción de la ruta.	-
route_type	Enum	✓	Modo de transporte de la ruta.	1
route_url	URL	✗	URL de una página asociada a la ruta.	https://www.crtm.es/4___1____.aspx
route_color	Color	✗	Color asociado a la ruta.	2DBEF0
route_text_color	Color	✗	Color del texto asociado a la ruta.	FFFFFF
route_sort_order	Número entero no negativo	✗	Número de la ruta asociado a un orden determinado.	-
continuous_pickup	Enum	✗	Posibilidad de subida a la ruta en cualquier punto del recorrido.	-
continuous_drop_off	Enum	✗	Posibilidad de bajada de la ruta en cualquier punto del recorrido.	-

Tabla 2.5: Campos del fichero *routes.txt* de la especificación GTFS Schedule

Datos AVL. Los datos AVL son datos de localización vehicular automatizada (AVL) registrados a intervalos regulares de tiempo sobre el geoposicionamiento de los vehículos de la red de transporte en un instante temporal determinado. Los sistemas AVL que generan

estos datos se sirven comúnmente de la tecnología GPS para llevar a cabo el rastreo de posición de los vehículos durante su recorrido en los viajes.

En la literatura [26, 27, 28, 29, 30, 31, 32, 33, 34] aparece como habitual la integración de estos datos con los correspondientes a transacciones de subida de pasajeros en las que solo se dispone de información temporal, sin localización, con el objetivo de inferir la ubicación concreta de la parada donde los pasajeros embarcan en el transporte público.

Datos APC. Los datos APC se corresponden con datos de sensores instalados en los diferentes medios de transporte público con los que proporcionar contadores de pasajeros automáticos (APC) en cada ubicación de la red de transporte. El hecho de conocer el número de pasajeros que pasan por cada parada de transporte puede resultar de gran interés de cara a propósitos de validación de las matrices OD de tránsito.

Datos de encuestas. Los datos de encuestas son datos históricos de patrones de tránsito recogidos a través de entrevistas con pasajeros del transporte público, ya sea en las propias instalaciones de la red o en sus domicilios, orientadas a conocer sus patrones de viaje a lo largo de la red de transporte público. Su uso principal se hace con fines de validación, para comprobar la semejanza entre los resultados obtenidos mediante los métodos de reconstrucción de viajes y las conclusiones deducidas de las encuestas. Hay que tener en cuenta que la utilización de este tipo de datos debe hacerse de forma rigurosa, pues circunstancias como una significativa diferencia temporal entre los datos de transacciones y de encuestas o aspectos demográficos diferentes entre las poblaciones de estimación y validación pueden llegar a hacer inconsistentes los resultados del estudio [27].

En general, el problema de reconstrucción de viajes de pasajeros a abordar es dependiente de la ciudad sobre la que se necesite aplicar, pues comúnmente cada ciudad presenta un conjunto de aspectos culturales propios y una topología de red de transporte que deben ser tenidos en cuenta a la hora de desarrollar un modelo de estimación de flujos, junto con sus asunciones asociadas, que se adapte a ella. Para poder tratar con este problema, se necesita conocer y, probablemente, estimar cierta información involucrada en la dinámica del tráfico de los pasajeros dentro de la red de transporte.

2.1.3. Desafíos

De acuerdo con los tipos de datos que tengamos disponibles, deberemos abordar varios desafíos que se presentan comúnmente a la hora de reconstruir los flujos de viaje de los pasajeros de una red de transporte público. En general, la tipología concreta del sistema AFC será la que determine la necesidad de abordar un desafío u otro.

Estimación de la parada de bajada. El problema de estimación de la parada de bajada (*alighting stop estimation*) surge ante la falta de transacciones de viaje que representen la bajada de los pasajeros en redes de transporte público que utilizan un sistema AFC de tipo *entry-only*. Para abordar la resolución de este problema se debe tener en cuenta

la sucesión de transacciones de subida al transporte público agrupadas por pasajero, de manera que se pueda inferir la información de bajada a partir de ellas. En este tipo de problema no se dispone de información temporal ni espacial sobre la bajada, por lo que los métodos que intenten realizar una estimación tendrán que asumir ciertas hipótesis y tener en cuenta algunas otras consideraciones importantes, de cara a proveer un resultado consistente.

Estimación de la parada de subida. El problema de estimación de la parada de subida (*boarding stop estimation*) se asemeja en parte al anterior, en tanto en cuanto no se dispone de toda la información necesaria sobre una parte relevante del viaje de los pasajeros. En este caso, las transacciones puede suponerse que han sido registradas mediante un sistema AFC de tipo *exit-only*, si únicamente se registran transacciones en la bajada de los pasajeros, o un sistema AFC de los tipos *entry-only* o *entry-exit*, en el caso de disponer de información parcial, comúnmente temporal, sobre la subida de los pasajeros en los medios de transporte público. La primera situación se trataría de forma equivalente al problema de estimación de la parada de bajada, donde se asumen ciertas hipótesis a priori coherentes. Mientras, en la segunda de las situaciones, que además es la más habitual, la localización espacial de las paradas de subida se estima a través de la integración de las transacciones con información horaria de viajes y/o de geoposicionamiento de vehículos.

Por otro lado, en la mayor parte de los casos también se necesita un mecanismo para la resolución de otro problema relativo al propósito de viaje de los pasajeros como es la detección de transbordos.

Detección de transbordos. La detección de transbordos (*transfer detection*) es una tarea que debe ser abordada cuando se requiere distinguir entre paradas de bajada acometidas para llevar a cabo un transbordo (ya sea entre distintos modos de transporte o entre vehículos de un mismo modo) y paradas de bajada correspondientes al destino final, donde el pasajero llega para llevar a cabo una actividad ajena al viaje. Percatarse de esta diferencia es clave para no determinar viajes incorrectos, que no representen el propósito real de los pasajeros.

La detección de estos transbordos se aborda habitualmente utilizando una serie de reglas básicas para diferenciar si se ha producido un transbordo o una actividad. En este caso, nos centraremos en la definición de las 3 reglas básicas que se vienen utilizando extensamente en la literatura:

- **Regla espacial**: Esta regla establece la distancia máxima de transbordo, o *maximum transfer distance* (MTD), que puede haber entre la parada de bajada de un segmento de viaje y la de subida del siguiente para considerar la ocurrencia de un transbordo. En caso de superarse este umbral de distancia entre paradas, se supondrá que el

pasajero se ha bajado del medio de transporte público para la realización de una actividad.

- **Regla temporal**: De forma semejante a la regla anterior pero aplicada en el ámbito temporal, esta regla establece el tiempo máximo de transbordo, o *maximum transfer time* (MTT), que puede transcurrir entre el instante de tiempo correspondiente a la bajada de un segmento de viaje y el correspondiente al de la subida del siguiente para considerar la ocurrencia de un transbordo. En caso de superarse este umbral de tiempo entre paradas, se supondrá que el pasajero se ha bajado del medio de transporte público para la realización de una actividad.
- **Regla de mismo sentido de ruta**: Esta regla establece que el pasajero ha bajado del transporte público para realizar una actividad siempre que 2 segmentos de viaje consecutivos sigan el mismo sentido de una misma ruta. En caso contrario, se deberán evaluar las demás reglas para determinar si realmente el pasajero ha bajado del transporte público para realizar una actividad o para continuar el viaje hacia su destino, es decir, realizando un transbordo.

Para determinar la ocurrencia de un transbordo entre segmentos de viaje se debe evaluar de forma completa este conjunto de reglas pues el incumplimiento de alguna de ellas, ya sea por superación de uno de los umbrales de distancia o tiempo o por el seguimiento de un mismo sentido de ruta, implicará directamente la consideración de una actividad entre paradas.

Una representación espacio-temporal de esta tarea de detección de transbordos puede verse en la Figura 2.5. En ella se representan 6 segmentos de viaje realizados por un pasajero durante un día de viaje, cuyo encadenamiento estará supeditado a la consideración de un transbordo o una actividad. Mientras las 4 primeras situaciones se asocian con actividades, por incumplirse alguna de las reglas de detección de transbordos (superar el umbral de MTD o MTT, o seguir un mismo sentido de ruta), la última situación sí cumple todas las condiciones requeridas para considerarse el transbordo y poder encadenar ambos segmentos de viaje consecutivos. Así, de acuerdo a estas reglas de transbordos, se consideraría la ocurrencia de un tranbordo entre la parada de bajada A_5 y la parada de subida siguiente B_6.

En nuestro caso, optaremos por el uso de unos determinados valores de parámetros para la detección de los posibles transbordos entre segmentos de viaje consecutivos de acuerdo con las 3 reglas básicas de distancia, tiempo y mismo sentido de ruta.

2.1.4. Métodos

De acuerdo con este conjunto de problemas descritos, en la literatura se proponen principalmente 3 aproximaciones diferenciadas para su resolución. Hay que tener en cuenta

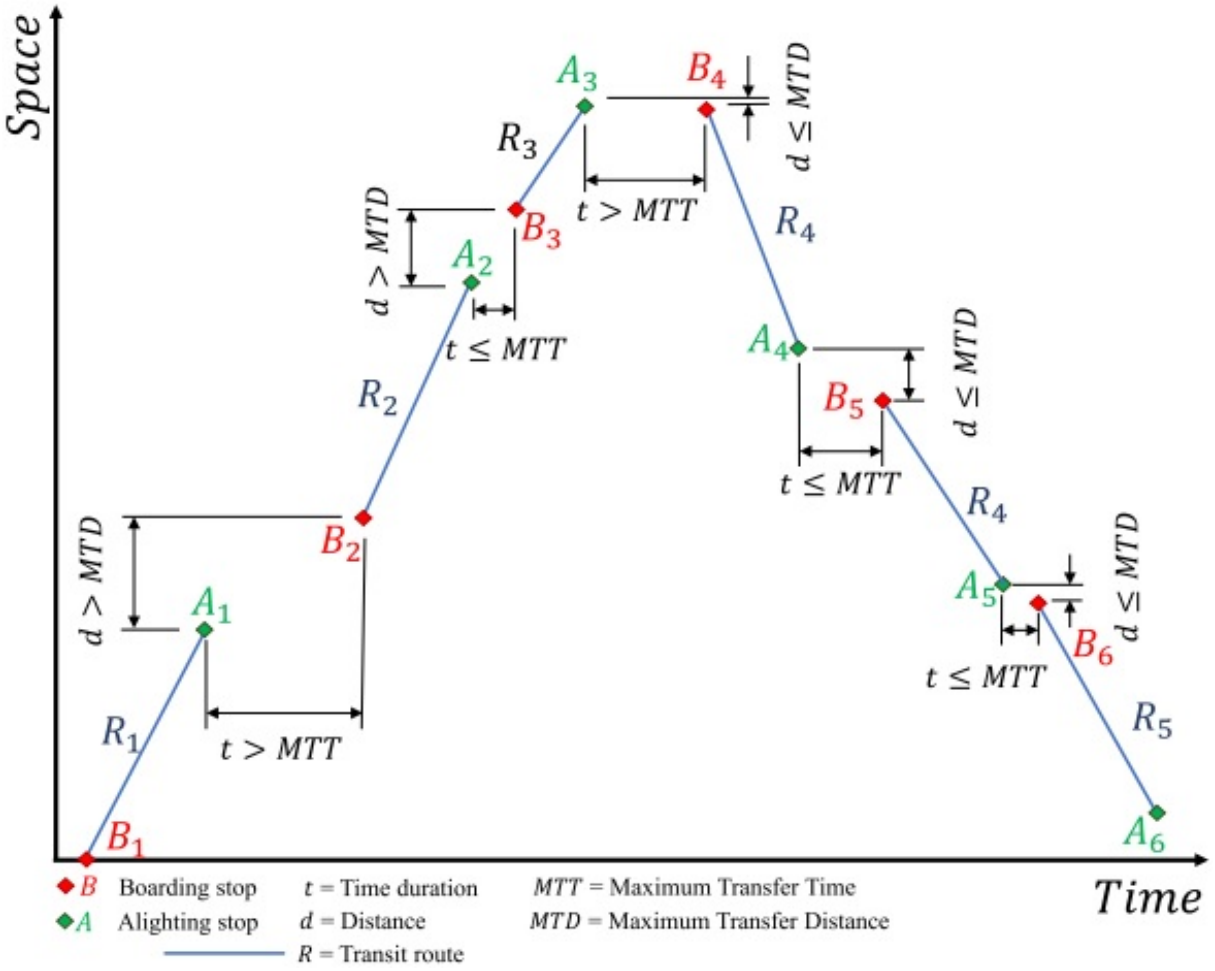

Figura 2.5: Representación espacio-temporal de la tarea de Detección de Transbordos [5].

que la cantidad y calidad de los datos de los que se disponga influenciará inevitablemente en los métodos de estimación más adecuados para resolver los problemas ya mencionados y, de hecho, condicionará la propia posibilidad de su uso efectivo.

Método *trip chaining*. El método *trip chaining* es un método estadístico basado en una serie de reglas y suposiciones que surge como mecanismo para tratar el problema de estimación de las paradas de bajada de pasajeros en una red de transporte [19]. El concepto de regla o, equivalentemente, suposición en el que se basa este método de estimación puede verse como una condición determinada a través de un proceso empírico, basado en la experiencia, con la cual se pretende establecer la forma en que se asumen ciertos hechos probados. La idea intuitiva detrás de este método es la deducida por su nombre, pues consiste en ir encadenando adecuadamente los segmentos de viaje (*trip legs*) de cada pasajero para conformar los viajes origen-destino que éste ha realizado, con el objetivo de llegar a un destino en el que llevar a cabo una determinada actividad.

Trip chaining es el método más utilizado en la literatura para abordar este problema, siendo muy intuitivo a la hora de su aplicación y aportando unos resultados convincentes en muchas redes de transporte publico, teniendo en cuenta que la precisión en la estimación siempre será dependiente de la cantidad y calidad de los datos de los que se dispone, así como del proceso de limpieza de datos (*data cleaning*) que se realice previo a la aplicación del método. Este método basa su funcionamiento en el uso de 2 suposiciones principales, cuya definición se corresponde con un proceso previo de revisión de los patrones de viaje que realizan la gran mayoría de los pasajeros en el transporte público:

- **Suposición de Continuidad**: Esta suposición considera la existencia de una determinada cercanía entre la parada de bajada de un segmento de viaje y la de subida del siguiente. De esta forma, se supone que dicha parada de bajada estará ubicada dentro de una zona de un determinado radio de distancia alrededor de la parada de subida del segmento siguiente. Este concepto de zona, en la que considerar las paradas de bajada candidatas para un segmento de viaje, se denominada comúnmente *Buffer Zone*. Así, se establece un determinado umbral, parametrizable de acuerdo con las características propias de la red de transporte en cuestión, a partir del cual poder tomar una decisión de estimación acerca de las paradas de bajada de los pasajeros que hayan sucedido de forma más probable en la realidad de los viajes.
- **Suposición de Simetría**: Esta suposición considera que la parada de bajada del último viaje del día realizado por un pasajero finaliza en la parada de subida del primero. Por lo tanto, supone que el origen del primer viaje del día y el destino del último coinciden, pues se considera que el pasajero vuelve a su domicilio. Esta suposición es comúnmente relajada en la literatura a través de (i) la suposición de una distancia de cercanía con respecto a la parada de origen del día actual o del siguiente, para no estimar necesariamente la misma parada de origen, y (ii) la consideración de días virtuales en los que los rangos de horas no coinciden con los habituales (de 00:00 a 23:59).

Un esquema del funcionamiento de este método aparece en la Figura 2.6. En ella podemos apreciar 3 segmentos reales de viaje caracterizados por sus respectivas paradas de subida y bajada representadas como B_i y A_i, respectivamente. Este método propone la definición de una *Buffer Zone* alrededor de cada parada de subida (B_i) para determinar la parada de bajada del segmento de viaje anterior (A_{i-1}). Para el primer segmento de viaje, se localiza su parada de bajada (A_1) dentro de la zona considerada alrededor de (B_2), mientras que para el segundo segmento de viaje vemos que su parada de bajada (A_2) no se ha podido determinar a través de la zona alrededor de (B_3). Por último, el tercer segmento de viaje del día termina (A_3) donde empezó el primero del día (B_1).

Otro ejemplo de aplicación del método *trip chaining* más visual y que representa el prototipo de viajes realizados por un pasajero durante un día laboral puede verse en la Figura 2.7. En ella se muestran 3 segmentos de viaje correspondientes a los movimientos

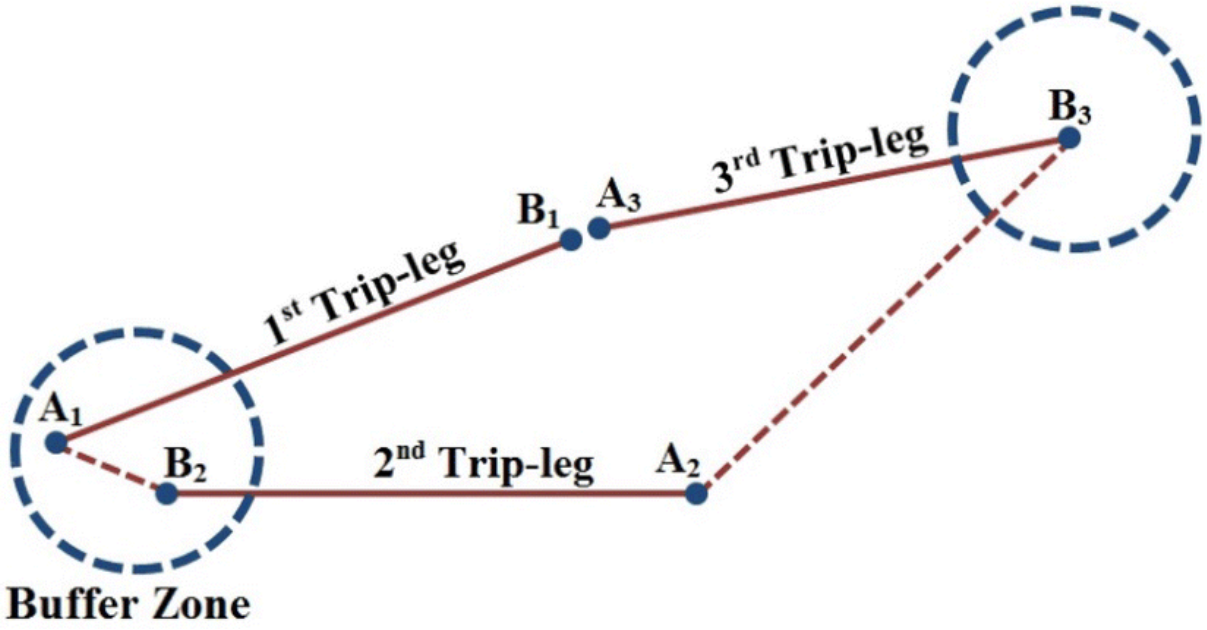

Figura 2.6: Esquema del funcionamiento de *trip chaining* [6].

del pasajero. El primer segmento describe el viaje entre su domicilio y su lugar de trabajo, el segundo entre su lugar de trabajo y el centro comercial y el tercero entre el centro comercial y su domicilio. De acuerdo con la suposición de continuidad del método *trip chaining*, se establece una distancia umbral máxima desde cada parada de subida (B_i) para encontrar la parada de bajada anterior (A_i) más cercana, considerando, según la suposición de simetría, la primera parada de subida del día (B_1) como la correspondiente a la última parada de bajada de ese día.

Finalmente, este método *trip chaining* se suele emplear junto con el conjunto básico de reglas (MTD, MTT y mismo sentido de ruta) utilizadas para detectar la ocurrencia de posibles transbordos entre segmentos de viaje.

Métodos probabilísticos. El uso de nuevos métodos para la estimación de la paradas de bajada de los pasajeros en el transporte público surge como medio para relajar la rigidez de los métodos basados en reglas, como *trip chaining*, con el objetivo de proveer una estimación más flexible que permita adaptarse mejor al comportamiento y hábitos de viaje de los pasajeros del transporte público. En concreto, los métodos de tipo probabilístico se basan en asignar una probabilidad a cada parada seleccionada como candidata, de acuerdo a una serie de condiciones como la distancia o el tiempo entre la parada de bajada y la subsiguiente de subida, la capacidad de pasajeros y el número de posibles transbordos de la propia parada o el tipo de uso de la zona (*land use*) alrededor de ella. Uno de estos métodos probabilísticos utilizados en la literatura es la asignación latente de Dirichlet (*latent Dirichlet allocation*) [7], el cual aparece representado en la Figura 2.8. En ella se

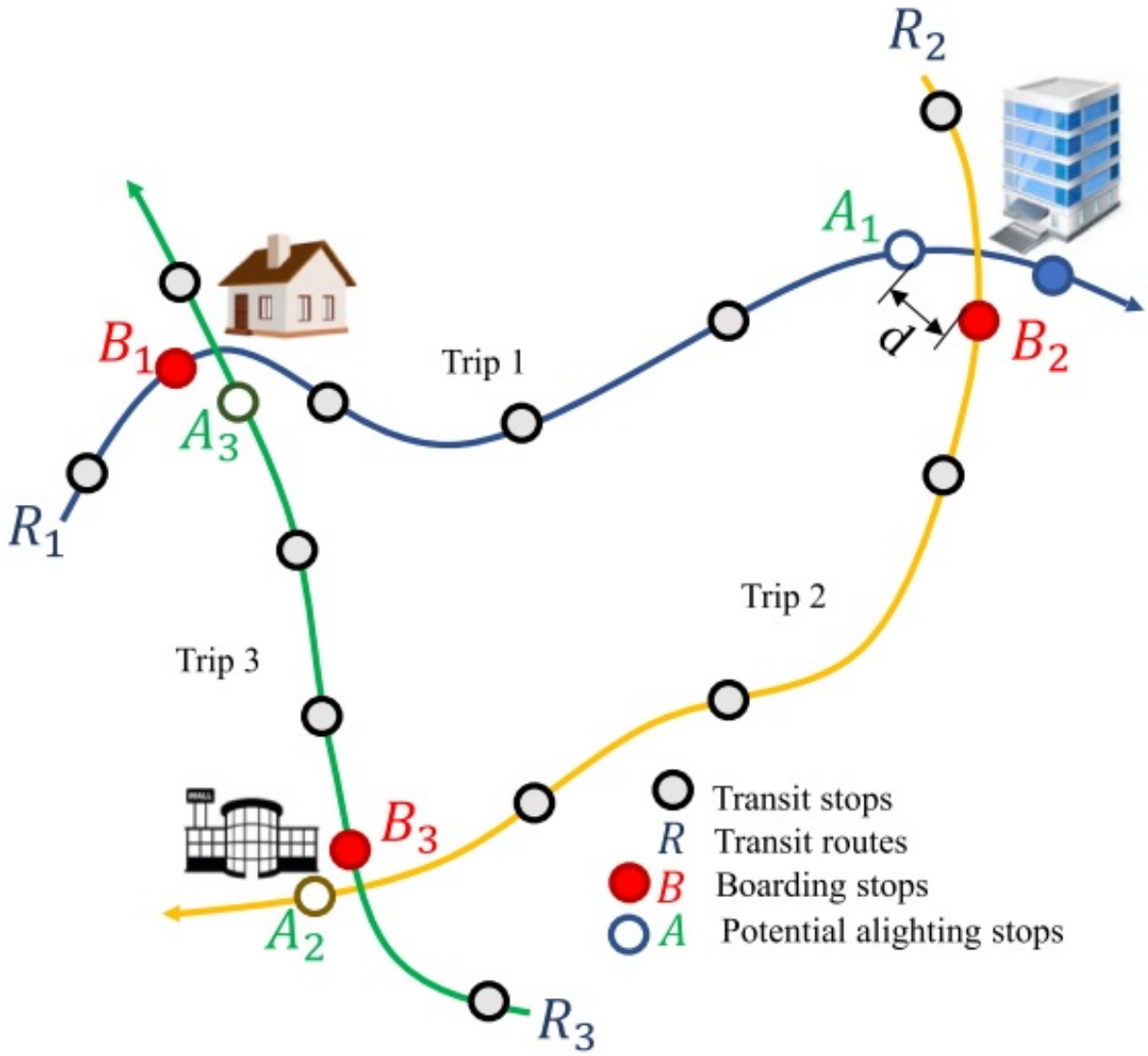

Figura 2.7: Representación gráfica del método *trip chaining* [5].

muestra el esquema de un modelo tridimensional de asignación latente de Dirichlet en el que se estiman, a través de una serie de cálculos, las distribuciones de probabilidad asociadas a las 3 variables que constituyen cada viaje de un pasajero. Estas variables son el tiempo de salida (w^t), la parada de salida (w^o) y la parada de bajada (w^d).

Métodos basados en *deep learning*. Gracias al auge del aprendizaje automático y a la necesidad de relajar las restrictivas suposiciones del método *trip chaining*, los métodos de *deep learning* basados en redes neuronales surgen con la idea de incorporan una mayor cantidad y variedad de información de entrada de cara a inferir las paradas de bajada de los pasajeros más probables en cada situación.

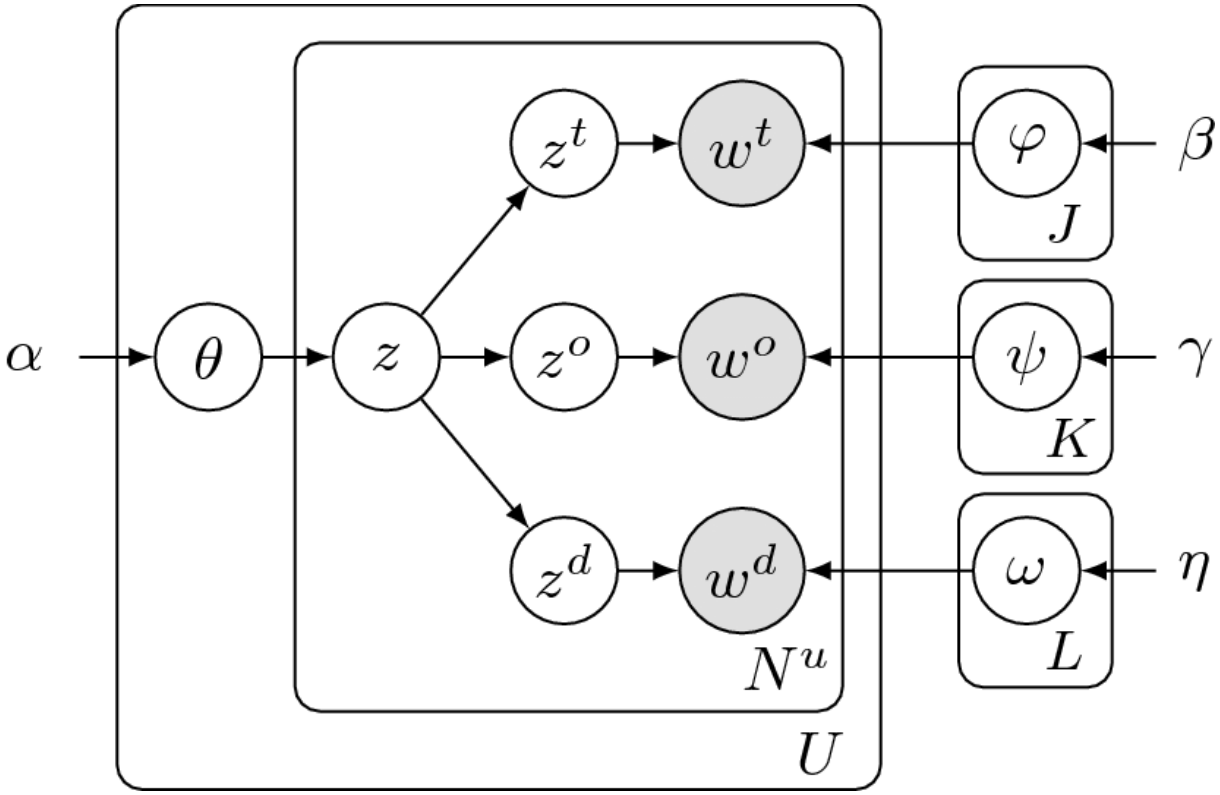

Figura 2.8: Representación gráfica del modelo probabilístico LDA (*latent Dirichlet allocation*) [7].

Este tipo de métodos suelen tener en cuenta multitud de variables de entrada, relativas tanto a las transacciones como a la zona en la que se encuentra la parada de subida considerada, para determinar el valor de las variables de salida de la red neuronal, es decir, las probabilidades asociadas a cada parada candidata de bajada del segmento de viaje. Una representación gráfica de este tipo de redes neuronales aparece en la Figura 2.9. En ella se presenta el esquema de una red neuronal profunda con una capa de entrada, otra de salida y dos ocultas intermedias. La capa de entrada (*input layer*) recibe las diferentes variables tanto relativas a las transacciones de viaje como a características urbanas de la zona *land use*. Por su parte, la capa de salida devuelve la probabilidad asociada a cada una de las 5 paradas candidatas para la estimación de la parada de bajada. En cuanto a las dos capas ocultas de la red neuronal, su funcionamiento se basa en el ajuste de unos pesos (*weights*) que van cambiando a medida que se entrena la red neuronal con nuevas muestras.

El mayor inconveniente de este tipo de métodos es la necesidad de disponer de grandes conjuntos de datos ya etiquetados, con información de la subida y bajada de los pasajeros, para el entrenamiento y validación del modelo de estimación. Este hecho dificulta su uso efectivo en las redes de transporte público que utilizan sistemas AFC de tipo *entry-only*

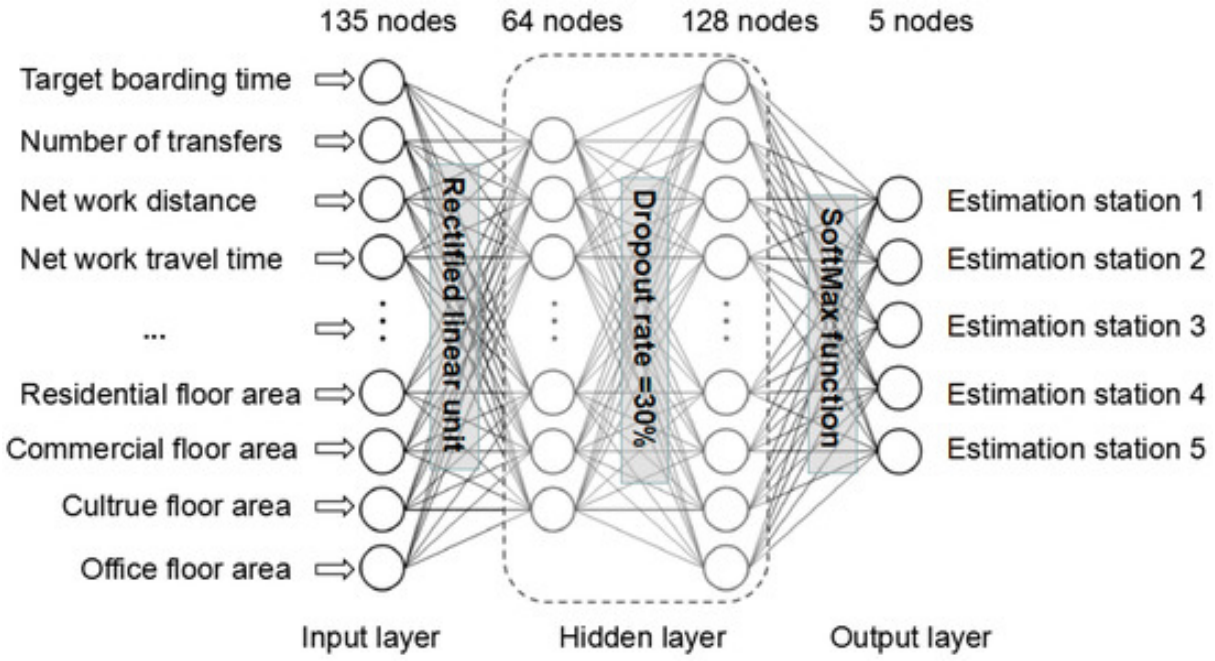

Figura 2.9: Representación de red neuronal profunda (*deep learning*) [8].

para el registro de las transacciones de viaje, como la de Madrid. Por tanto, el uso de este tipo de técnicas de aprendizaje automático se reserva para las redes de transporte público que utilicen sistemas AFC de tipo *entry-exit*.

2.2. Red de Transporte Público de Madrid

Una red de transporte público es un conjunto integrado de servicios de movilidad de pasajeros que constituyen una infraestructura para la interconexión de diferentes zonas de tránsito de un determinado área a través del uso de vehículos de transporte con el objetivo de favorecer la movilidad de las personas por las diferentes ubicaciones de la red. Es frecuente que estas redes de transporte público integradas abarquen las diferentes zonas pertenecientes a un mismo área de administración territorial, ya sea un municipio, concejo, región, país o, incluso, una unión de varias de ellas bajo un determinado acuerdo [35]. Así, estas redes de transporte público habilitan la posibilidad a los ciudadanos de moverse libremente, dentro de unos horarios de servicio convenidos, a lo largo de toda la red a través de los medios de transporte puestos a su disposición por las empresas operadoras encargadas del servicio.

En el caso de Madrid, su red de transporte público es una red intermodal que cuenta con diversos modos de transporte enlazados a través de paradas de correspondencia para

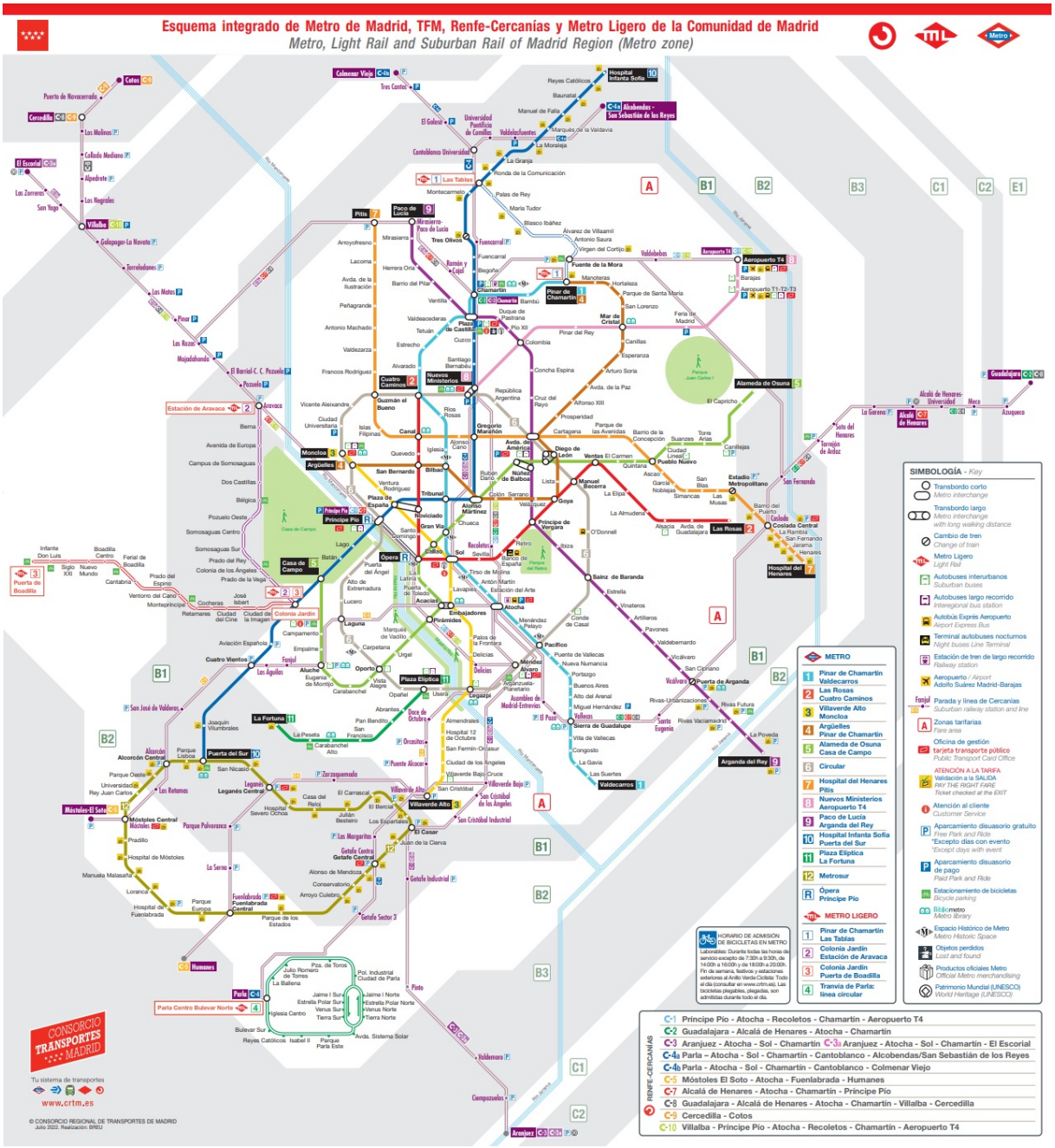

Figura 2.10: Topología Parcial de la Red de Transporte Público de Madrid [9].

conformar un entramado de movilidad integrado en el que un mismo viaje realizado por un pasajero pueda pasar por los diferentes modos de transporte constituyentes. Un mapa con la topología parcial de la red de transporte público de Madrid a fecha de julio de 2022 puede consultarse en la Figura 2.10. Se puede comprobar que las estaciones de esta red de transporte público se encuentran ubicadas en diferentes zonas concéntricas alrededor de la

ciudad de Madrid. Estas zonas, a lo largo de las cuales se dispersan las estaciones, son referidas como coronas tarifarias, siendo su distribución tal y como aparece en el mapa de la Figura 2.11.

Figura 2.11: Coronas tarifarias de la red de transporte público de Madrid [10].

La gestión de toda esta red de transporte público de Madrid es llevada a cabo por el Consorcio Regional de Transportes de Madrid, "*un organismo público fundado en el año 1985 que concentra las competencias sobre transporte público regular de viajeros de la Comunidad de Madrid y los Ayuntamientos que se adhieran al mismo*"[36]. Este organismo se encarga de la gestión de las infraestructuras, recursos y servicios, así como de la tramitación de autorizaciones y concesiones a agencias operadoras externas, de cara a ofrecer una red de transporte público debidamente cohesionada con, además, un sistema tarifario integrado. Este sistema tarifario posibilita una entrada sencilla a cada uno de los modos de transporte de cara a facilitar al pasajero sus viajes por la red. Una visión del marco institucional del Consorcio Regional de Transportes de Madrid puede servirnos para observar los modos de transporte que son gestionados de forma integrada por este organismo (véase la Figura 2.12).

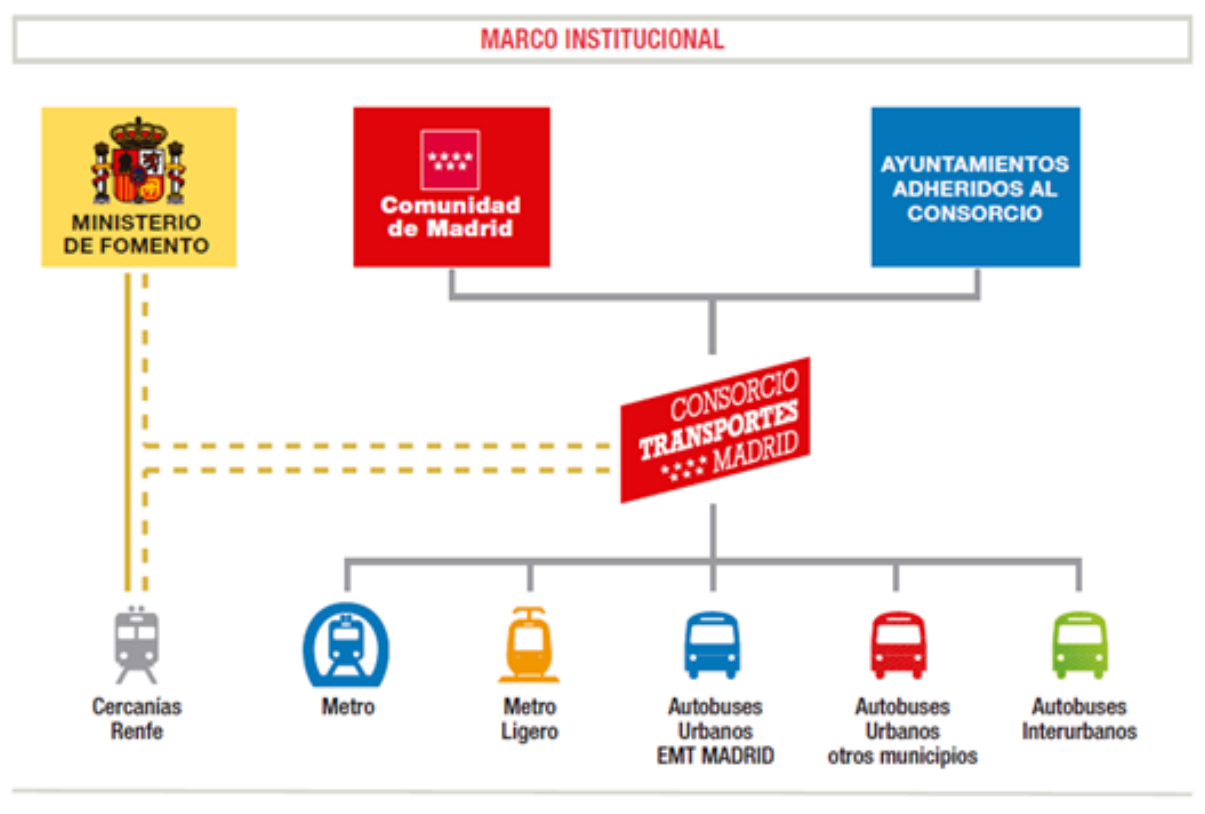

Figura 2.12: Marco Institucional del Consorcio Regional de Transportes de Madrid [11].

En cuanto a los modos de transporte que componen esta red intermodal de transporte público de Madrid, éstos pueden ser agrupados en torno a 4 tipos generales [37]:

Metro. Dentro de este modo de transporte se puede hacer una diferenciación entre los 2 siguientes tipos:

- **Metro de Madrid**: El Metro de Madrid es el modo suburbano por excelencia de la red de transporte público de Madrid y está compuesto por 12 líneas más 1 ramal, lo que supone una longitud total de 287 km y 2394 vehículos ofreciendo servicio. La red en su conjunto acumula un total de 238 paradas que son utilizadas para la subida y bajada de los pasajeros a la red de Metro.
- **Metro Ligero de Madrid**: El Metro Ligero de Madrid es una red de tren ligero constituida por un total de 4 líneas y 52 paradas que conecta el área metropolitana de la Comunidad de Madrid con la ciudad de Madrid, abarcando un total de 35,4 km de vía.

Autobuses Urbanos. Igualmente, dentro de ese modo de transporte se puede hacer una diferenciación en los 2 siguientes tipos:

- **Autobuses Urbanos EMT de Madrid**: Los Autobuses Urbanos EMT de Madrid ofrecen un servicio de autobuses en la ciudad de Madrid que cuenta con 203 líneas y más de 10172 paradas, lo que supone una longitud total 3562 km abarcados y 1903 vehículos ofreciendo servicio.
- **Autobuses Urbanos de otros municipios**: Los Autobuses Urbanos de otros municipios ofrecen un total 109 líneas de autobús que operan dentro de otros municipios de la Comunidad de Madrid. En total, este modo de transporte realiza más de 6155 servicios diarios.

Autobuses Interurbanos. Los Autobuses Interurbanos conforman una red de transporte de autobuses intermunicipal compuesta por 324 líneas que realizan más de 20000 servicios al día abarcando un total de 19065 km de extensión.

Cercanías. Cercanías es una red ferroviaria que opera en la Comunidad de Madrid y cuenta con 9 líneas y 92 estaciones de tren, abarcando un total de 391 km de vía.

Para la interconexión de todos los modos de transporte, de cara a ofrecer una red integrada, se establecen paradas de correspondencia en las cuales los pasajeros pueden hacer transbordo sin necesidad de salir de las infraestructuras de la red de transporte. De esta forma, una parada que forma parte de una ruta en uno de los modos de transporte puede servir como punto de transbordo para el acceso a otros modos distintos o para el cambio a otras rutas en el mismo modo.

A la hora de registrar una transacción de viaje en los terminales del sistema AFC instalados a lo largo de la red de transporte se considera el uso de 2 tipos de tarjetas de transporte diferenciadas:

- **Tarjeta de transporte público personal**: es una tarjeta de transporte de tipo personal, que incluye los datos demográficos de la persona a la que pertenece.
- **Tarjeta de transporte público multi**: es una tarjeta de transporte no personal, con una duración de 10 años que no recoge información alguna sobre su poseedor.

Además, se ofrecen diversas opciones tarifarias que pueden ser cargadas dentro de los tipos de tarjetas anteriores, de acuerdo con las necesidades requeridas por los pasajeros.

- **Billetes**: son títulos tarifarios básicos que pueden ser cargados en los 2 tipos de tarjeta de transporte anteriores y que están designados comúnmente para los pasajeros

ocasionales que requieren viajar en el transporte público un número reducido de veces. Dentro de esta opción, las 2 modalidades básicas son: el billete sencillo (válido para 1 viaje) y el billete de 10 viajes.

- **Abonos**: son títulos tarifarios de carácter personal que puede ser cargados únicamente en las Tarjetas de Transporte Público Personal y que se caracterizan por ser económicamente más rentables para los pasajeros que realizan un número de viajes considerable. Entre las modalidades de abono que se pueden adquirir según las necesidades de transporte de los pasajeros destacan: el abono anual, el abono 30 días y los abonos de familia numerosa y personas con discapacidad. Asimismo, se definen 3 perfiles de usuario principales a la hora de acceder a los tipos de abono anual y 30 días: normal (entre 26 y 64 años), joven (menores de 26 años) y tercera edad (a partir de 65 años).

En nuestro proyecto, las transacciones de viaje disponibles se corresponden con las realizadas a través de tarjetas de transporte público personales utilizando las variadas opciones tarifarias ofrecidas por el CRTM. Este tipo de tarjetas tienen asociado un identificador único que habilita el seguimiento de los trayectos de viaje consecutivos realizados por los pasajeros, necesario para la aplicación del método de estimación *trip chaining* que se propone.

2.2.1. Datos disponibles

En este proyecto disponemos de 2 tipos de datos diferenciados para llevar a cabo la estimación de las paradas de bajada de los pasajeros y así poder reconstruir sus flujos de viaje y componer las matrices OD de tránsito.

Transacciones de viaje. Los datos de transacciones de viaje son registros recopilados por medio de terminales instalados en las diferentes ubicaciones de la red de transporte público para registrar la subida de los pasajeros a los medios de transporte. La mayor parte de los terminales del sistema AFC instalados son utilizados únicamente para registrar la información de subida/entrada de los pasajeros a la red de transporte, lo que implica que, en general, el sistema AFC sea de tipo *entry-only*. Este hecho se traduce en la ausencia de información sobre la bajada/salida del pasajero del transporte público. La excepción a esta regla aparece en los viajes realizados en metro, metro ligero y cercanías, donde las transacciones de viaje correspondientes incluyen información tanto de la subida como de la bajada de los pasajeros.

Además, hay que tener en cuenta que en ciertos modos de transporte, como el metro o cercanías, el registro de las transacciones se hace a la entrada de las estaciones sin poder saber, por tanto, la ruta ni el sentido de ruta que el pasajero seguirá una vez haya accedido

a su interior. Asimismo, la información sobre transbordos en este tipo de modos será a priori transparente para nosotros.

En nuestro caso, un conjunto completo de datos de transacciones ha sido puesto a nuestra disposición por el Consorcio Regional de Transportes de Madrid para la realización de este trabajo (ver Sección 4.2). En este conjunto de datos no se incluyen las transacciones de viaje realizadas en el modo de transporte autobuses urbanos de otros municipios, ya que los movimientos de los pasajeros se limitan a los propios municipios y, por ello, no hay posibilidad de interconexión directa con otros modos de transporte que operen en municipios diferentes. Así, el contexto de aplicación queda delimitado a únicamente las transacciones de viaje registradas en los modos de transporte que permiten moverse por la ciudad de Madrid o entre ésta y otros municipios de la Comunidad de Madrid.

Por otro lado, en las Tablas 2.6 y 2.7 se enumeran los perfiles de usuario y títulos tarifarios que están presentes en las transacciones de viaje puestas a nuestra disposición. Así, se puede comprobar la gran variedad de tipos de perfiles y opciones tarifarias que existen en la gestión de la red de transporte público de Madrid.

Denominación Perfil de Usuario
Normal
3ª Edad
Joven
Infantil
Tarjeta Azul
Turístico Normal
Turístico Infantil
Programa Activación y Empleo
Pensionista Valdemoro
Discapacitado 33 % Valdemoro

Tabla 2.6: Perfiles de Usuario representados en las transacciones de viaje

Datos generales de la red de transporte público de Madrid. Estos datos son relativos al funcionamiento general de la red de transporte público de Madrid y son ofrecidos públicamente por el Consorcio Regional de Transportes de Madrid a través de su Portal de Datos Abiertos [38].

Entre toda la información aportada en este Portal de Datos Abiertos, se destacan para la realización de este proyecto los datos estáticos de la red que definen su topología y proporcionan información de utilidad para la caracterización de las rutas de viaje y las paradas de la red de transporte (ver Sección 4.2).

Algunos conjuntos de datos incluidos dentro del Portal de Datos Abiertos del Consorcio Regional de Transportes de Madrid que son de potencial utilidad para su integración

Tipo	Denominación Título de Transporte
Billete Sencillo	Metro (5, 6, 7, 8, 9 y 10 estaciones), MetroNorte, MetroSur, MetroEste, MLO, TFM, Combinado, Intrazonal (A, B1, B2, B3, C1 y C2), Interzonal (A, B1, B2, B3, C1, C2, E1 y E2), Urbano Rivas, Corte Servicio Metro.
Billete 10 Viajes	MetroNorte, MetroSur, MetroEste, MLO, TFM, EMT, Metrobús, Combinado.
Abono 30 Días	Tercera Edad, Joven Tarifa Plana, Tarjeta Azul, Programa Activación y Empleo, Intrazonal (A, B1, B2, B3, C1, C2, E1 y E2), Interzonal (A, B1, B2, B3, C1, C2, E1 y E2).
Abono Anual	Tercera Edad, Joven Tarifa Plana, Intrazonal (A, B1, B2, B3, C1, C2, E1 y E2), Interzonal (A, B1, B2, B3, C1, C2 y E1).
Abono Turístico	Zona A (1, 2, 3, 4, 5 y 7 días) y Zona T (1, 2, 3, 4, 5 y 7 días).
Abono Autobuses Urbanos	Intrazonal (B1, B2, B3, C1 y C2) y Rivas.
Abono Autobuses Interurbanos	Interzonal (A, B1, B2, B3, C1 y C2).
Abono Valdemoro	Pase Municipal Valdemoro.
Abono Infantil	Tarjeta Transporte Público Infantil.
Abono Empleados	Agente Metro y Mantenimiento Metro.

Tabla 2.7: Títulos de Transporte representados en las transacciones de viaje

con datos de transacciones de cara a la resolución de diferentes problemas de transporte urbano inteligente aparecen descritos en la Tabla 2.8.

Para obtener más detalles acerca de estos conjuntos de datos se recomienda visitar el sitio web del Portal de Datos Abiertos del Consorcio Regional de Transportes de Madrid (https://data-crtm.opendata.arcgis.com/).

La integración conjunta de los datos de transacciones con datos generales de la red de transporte público de Madrid será un aspecto crucial en el proceso ETL con el objetivo de ampliar el conocimiento acerca de los viajes realizados por los pasajeros.

2.3. Estado del Arte

A continuación, se presentará una recopilación con algunos de los estudios más destacados propuestos en la literatura para resolver el problema abordado en nuestro proyecto, es decir, la estimación de las paradas de bajada de segmentos de viaje en el transporte público, o bien, la construcción de las matrices OD de tránsito resultantes a partir de esos

Tipo de Información	**Descripción**	**Conjuntos de Datos**
Estaciones	Incluye información general sobre las estaciones de la red de transporte público.	▪ M4 Estaciones ▪ M10 Estaciones ▪ M5 Estaciones ▪ M6 Estaciones ▪ M8 Estaciones
Tramos por Ruta	Contiene información sobre los tramos que componen las rutas ofrecidas en el transporte público.	▪ M4 Tramos ▪ M10 Tramos ▪ M5 Tramos ▪ M6 Tramos ▪ M8 Tramos
Paradas por Ruta	Contiene información sobre las paradas por las que transcurren las rutas ofrecidas en el transporte público.	▪ M4 ParadasPorItinerario ▪ M10 ParadasPorItinerario ▪ M5 ParadasPorItinerario ▪ M6 ParadasPorItinerario ▪ M8 ParadasPorItinerario
Feeds GTFS	Incluye información general de la red de transporte público según la especificación GTFS.	▪ GTFS Red de Metro ▪ GTFS Red de Metro Ligero ▪ GTFS Red de Cercanías ▪ GTFS Red de EMT ▪ GTFS Red de Autobuses Interurbanos

Tabla 2.8: Conjuntos de datos interesantes del Portal de Datos Abiertos del CRTM.

flujos de tráfico determinados. Se llevará a cabo una revisión de cada uno de los estudios considerados para resumir su forma de trabajo y la solución adoptada para abordar la determinación de los flujos de pasajeros dentro de la red de transporte sobre la que trabajan. En este caso, cada estudio revisado propone una solución distinta adaptada y con matices acordes al contexto de trabajo sobre el que se aplica.

La motivación de mostrar una revisión detallada del estado del arte en torno al problema considerado radica en conocer qué han realizado ya otros autores para tratar la resolución del problema en cuestión para, así, tener una visión global sobre las posibles opciones a nuestro alcance a la hora de conformar nuestra propia solución. Seguidamente, se llevará a cabo una comparación entre las propuestas estudiadas y la de este proyecto, en la que se destacarán las principales características de cada una de ellas. Esta comparación se presentará en formato tabular para facilitar un rápido y visual cotejo entre las propuestas expuestas. Finalmente, se aportarán una serie de conclusiones deducidas a partir de la comparación y estudio individual realizado sobre cada una de las propuestas del estado del arte.

A continuación, se presenta un breve resumen del funcionamiento y peculiaridades consideradas en cada una de las propuestas analizadas, en orden cronológico por su fecha de publicación.

Barry et al. 2002 [19]

Este primer estudio trabaja con datos correspondientes a transacciones de viaje realizadas en el metro de Nueva York, Estados Unidos, en las que solo se tiene información sobre la localización y el tiempo de cada subida de un pasajero en las diferentes estaciones de metro. Por ello, al no tener información alguna sobre las estaciones de bajada de los pasajeros, se tratará de un sistema de recolección (AFC) de tipo *entry-only*.

Así, este estudio fue el primero en proponer el método *trip chaining* para estimar la localización de la parada de bajada de cada segmento de viaje (*trip leg*) realizado por un pasajero, representado por su tarjeta personal de transporte (*smartcard*). Este método *trip chaining* pretende la unión de los segmentos de viaje realizados en el transporte público por un pasajero en un día concreto, con el objetivo de conformar la ruta de viaje completa seguida por el pasajero desde la primera subida del día en el transporte público hasta la última.

Inicialmente, este método se propuso con 2 suposiciones base de partida, que serían posteriormente relajadas y adaptadas al entorno de trabajo en los siguientes estudios realizados en este tema. Estas suposiciones mencionadas son las siguientes:

- **Suposición de continuidad**: asume que los segmentos de viajes recorridos por un pasajero empiezan en una ubicación próxima a donde terminan los anteriores. De esta forma, la parada de bajada de un segmento de viaje se encontrará cerca, dentro de un determinado umbral de distancia máxima, de la parada de subida en la que el pasajero se embarca para su siguiente segmento de viaje. Esta circunstancia se deduce del hecho de que un gran porcentaje de los pasajeros empieza el siguiente segmento de viaje cerca del punto de finalización del anterior.
- **Suposición de simetría**: asume que el último segmento de viaje de un pasajero termina en la misma ubicación donde empezó el primer segmento de viaje del día.

> De esta forma, se supone que las paradas de bajada del último segmento de viaje del día y de subida del primer segmento son la misma. Esta circunstancia se deduce del hecho de que la gran mayoría de los pasajeros vuelven a su domicilio al final del día.

Tengamos en cuenta que este estudio considera que un pasajero solo utiliza el metro como medio de transporte, no pudiéndose inferir la parada de bajada en aquellos casos donde utilicen otros medios de transporte distintos. En esta situación, tiene sentido establecer una distancia máxima andando (*walking distance*) para reducir el número de posibles estaciones candidatas en las que el pasajero es susceptible de haberse bajado en el segmento de viaje considerado.

Por lo tanto, la suposición de continuidad facilita la estimación de la parada de bajada de un segmento de viaje al considerar solo las paradas que se sitúan dentro de la distancia umbral máxima desde la parada de subida del siguiente segmento de viaje y que, además, se encuentren en la ruta concreta seguida en el segmento de viaje actual. Por otro lado, la suposición de simetría diaria cobra sentido al suponer que un pasajero saldrá de su casa al inicio del día y regresará a la misma al final del día.

Empleando este método sobre transacciones de la red de metro de Nueva York, este estudio reportó una precisión del 90 % en la estimación de las paradas de bajada de los pasajeros en cada segmento de viaje recorrido. Este valor de precisión en la estimación se obtuvo al hacer una validación del método de estimación de las paradas de bajada empleando datos oficiales del New York Metropolitan Transportation Council (NYMTC) asociados a viajes diarios en el metro de 250 tarjetas de pasajeros.

Trépanier et al. 2007 [39]

Este segundo estudio pretende la estimación de la localización y tiempo asociados a las paradas de bajada de los segmentos de viaje realizados por pasajeros en una red de autobuses de la ciudad de Gatineau, Quebec en Canadá. Los datos disponibles en esta propuesta se corresponden con transacciones recolectadas a través de un sistema AFC de tipo *entry-only*, es decir, en el que los pasajeros solo registran su entrada en los autobuses y no su salida. Por tanto, solo dispone de la información temporal y espacial relativa a las parada de subida de los pasajeros en los autobuses.

Al igual que en [19], se propone el uso del método *trip chaining* para la estimación de la información de bajada del transporte público, aunque es el primer estudio que propone relajar la suposición de simetría diaria considerada en la versión original del método. En este caso, propone considerar que la primera parada de subida del día y la última parada de bajada del día se encuentren dentro de un determinado radio de distancia y no restringiéndose a la consideración de la misma parada que se proponía en el método original. Esta relajación se aplicó utilizando una distancia umbral máxima de 2000 metros

para mejorar la estimación de la parada de bajada en una red de transporte público compuesta por rutas de buses, con una topología más compleja que una red de metro.

Este estudio propuso analizar datos de transacciones de un mes completo para abordar uno de los principales inconvenientes del método *trip chaining*, como es la estimación de la parada de bajada para tarjetas de transporte (pasajeros) con una única transacción asociada en un día natural. De esta forma, el estudio proponía buscar en otros días del mes transacciones similares, con hora cercana y misma ruta que la de la transacción única, para tratar de estimar correctamente la parada de bajada del pasajero.

Para la validación del método, este estudio utilizó 2 conjuntos de datos de transacciones en la red de transporte de Gatineau, correspondientes a los meses de julio y octubre de 2003, aportados por la autoridad de tránsito de la región de Outaouais (STO). Los resultados reportados por el estudio muestran una precisión global en la estimación de la parada de bajada de los autobuses del 66 %, incrementando este valor de precisión hasta el 80 % de paradas de bajada correctamente inferidas durante las horas de demanda punta. Asimismo, en lo relativo a las transacciones únicas realizadas en un día, los resultados aseveran una precisión del 68 % en la estimación de sus paradas de bajada utilizando el método *trip chaining* adaptado.

Zhao et al. 2007 [26]

Este estudio también utiliza el método determinista *trip chaining* para la estimación de la matriz OD de viajes de tren en una red bimodal, formada por rutas de autobús y líneas de tren, de la ciudad de Chicago, Estados Unidos. En este caso, la peculiaridad del estudio radica en la inferencia de la información de origen de los segmentos de viaje en autobús a partir de la integración del conjunto de datos de transacciones, recolectadas por medio de un sistema AFC de tipo *entry-only*, con un conjunto de datos AVL (*Automatic Vehicle Location*) que registra la localización de los vehículos a intervalos regulares de tiempo. De esta forma, al contar las transacciones con su marca temporal correspondiente, se puede determinar el número de autobús y la ruta asociada, cruzando ambos conjuntos de datos.

Este estudio también trata la detección de transbordos entre líneas de tren y entre rutas de autobus y líneas de tren, con el objetivo de determinar los segmentos de viaje de un pasajero que deben encadenarse (por considerarse fruto de un mismo propósito de viaje), para así tener una información correcta del origen y destino de los recorridos de viaje de cada pasajero, que serán plasmados en la matriz OD resultante.

En esta propuesta, se utiliza el valor de 400 metros para el parámetro MTD (*maximum transfer distance*), con el que determinar si la parada de bajada de un segmento de viaje y la parada de subida del siguiente están lo suficientemente cercanas como para considerar a este siguiente segmento como parte del actual, es decir, establecer que el pasajero ha hecho un transbordo entre medios de transporte.

En el estudio se plantea una validación exógena, empleando datos de matrices OD recopilados en encuestas del Chicago Transit Authority (CTA) para validar las matrices OD obtenidas con el método. En este caso, los resultados del estudio arrojan una precisión del 90 % en la estimación de la matriz OD construida.

Farzin 2008 [27]

Este estudio propone la inferencia de la matriz OD para la red de autobuses de la ciudad de Sao Paulo, Brasil, mediante el método *trip chaining*, utilizando transacciones recopiladas de un sistema AFC de tipo *entry-only*. El hecho de tener solamente el tiempo de subida al autobús implica la necesidad de integrar las transacciones AFC con un conjunto de datos AVL, sistema con el que se encuentra equipado cada autobús, para determinar la localización de la parada de subida. En este caso, la disponibilidad en ambos conjuntos de datos del identificador del autobús y el instante de tiempo posibilitan la integración de ambos para deducir la ubicación actual de la parada de autobús. Con objeto de atenuar los errores de medición típicos en los sistemas GPS asociados a los datos AVL, se asume un radio de distancia umbral de 110 metros para hacer corresponder la parada de subida a estimar.

Esta propuesta considera la versión original del método *trip chaining*, asumiendo que la parada de bajada del último segmento de viaje del día coincide exactamente con la parada de subida del primer segmento de viaje del día. La validación, de tipo exógeno, en este estudio evalúa las diferencias entre la matriz OD estimada y la recopilada a partir de encuestas realizadas en los hogares de los pasajeros, observándose la presencia de diferencias significativas en cuanto a los destinos y patrones de viaje seguidos por los pasajeros encuestados. No obstante, la diferencia notable de años entre las transacciones con las que se estima la matriz OD (del año 2006) y las encuestas realizadas en los hogares a partir de las que se construye la considerara como matriz verdadera (del año 1997) puede hacer inconsistente la validación acometida, pues el comportamiento de los viajeros, la planificación de los recursos disponibles o, incluso, la topología de la red de transporte público pueden haber cambiado drásticamente a lo largo del tiempo.

Barry et al. 2009 [40]

Este estudio plantea la estimación mediante el método *trip chaining* de una matriz OD zonal para la red de transporte multimodal, compuesta por rutas de autobús y líneas de metro, de Nueva York, Estados Unidos. Los datos utilizados en el método propuesto se corresponden con transacciones provenientes de un sistema AFC de tipo *entry-only*, asociado a subidas de pasajeros en paradas de autobús y estaciones de metro, recogidas durante 2 semanas.

Una de las motivaciones de este nuevo estudio fue reducir el número de pasajeros con una única transacción durante el día, al observar que los patrones de viaje de ciertos pasajeros se prolongaban más allá de las 00:00 horas, en las que se considera el inicio de un nuevo día. Así, se propuso la definición de un día virtual que empezara a las 3:00 horas y terminara a las 2:59 del día siguiente, para intentar reducir el número de transacciones únicas en el nuevo rango de día virtual considerado. Nótese que este rango es muy dependiente de los usos y costumbres del lugar donde se pretenda utilizar el modelo de estimación, así como del tamaño de ciudad donde se aplique, pues en ciudades de tamaño pequeño/medio es poco común que el transporte público siga funcionando por la noche.

Al igual que en [19], se asumen las suposiciones del método *trip chaining* originalmente planteado, haciendo coincididir la parada de bajada del último segmento de viaje del día virtual con la parada de subida del primer segmento de viaje del día virtual habilitado. Por otro lado, a la hora de detectar los transbordos entre medios de transporte se utiliza un tiempo umbral máximo de 18 minutos, entre el tiempo de bajada del segmento actual y el tiempo de registro de la transacción en la subsiguiente parada de subida, a partir del cual se pasará a considerar que el pasajero se ha bajado del transporte público para realizar una actividad y, por lo tanto, no se debe encadenar el segmento de viaje siguiente al actual (*unlinked trip*).

La validación, de tipo exógeno, del método propuesto en el estudio se hizo utilizando contadores de entradas y salida de pasajeros en el caso de las estaciones de metro y datos de carga media y contadores de subida y bajada en el caso de las paradas de autobús. Sin embargo, para los autobuses únicamente se emplearon datos asociados a 2 rutas, lo cual no permite llevar a cabo una validación lo suficientemente robusta. Finalmente, los resultados de evaluación mostraron una precisión del 90 % en la estimación del paradas de bajada de los pasajeros.

Seaborn et al. 2009 [28]

Este estudio hace uso del método *trip chaining* para estimar la matriz OD de una red de transporte multimodal compuesta por rutas de autobús y líneas de metro de Londres, Reino Unido. Por un lado, para este estudio se cuenta con la información tanto de subida como de bajada asociada a los segmentos de viaje realizados por los pasajeros dentro de la red de metro, es decir, resultante de transacciones recolectadas de un sistema AFC de tipo *entry-exit*. Por otro lado, asociado a los segmentos de viaje realizados en autobús solo se dispone del instante de tiempo en el que el pasajero se embarcó en cada autobús. Por ello, para inferir la ubicación concreta de la parada de subida a un autobús se necesita integrar el conjunto de datos de transacciones AFC con un conjunto de datos AVL de rastreo de geoposicionamiento instalado en cada autobús.

En cuanto a la detección de transbordos, esta propuesta considera el uso de distancias máximas de transbordo (MTD) diferentes según el tipo de cambio entre modos de transporte

dentro de la red multimodal, al considerar que realmente son distintas las distancias medias a recorrer por un pasajero para moverse de un medio de transporte a otro. Por ejemplo, comúnmente no se necesitará recorrer la misma distancia para moverse entre 2 líneas de autobús que para moverse del metro a una línea de autobús. Igualmente, considera rangos de tiempo diferenciados para el tiempo máximo de transbordo (MTT) que varían según el cambio entre los modos de transporte de la red. En este caso, se utilizan los rangos 15-25 min, 30-50 min y 40-60 min para evaluar las combinaciones de transbordos entre metro-bus, bus-metro y bus-bus, respectivamente. Para los transbordos realizados entre líneas de metro no se registra transacción alguna, por lo que no se podría utilizar este método para detectarlos.

Este estudio sólo plantea un proceso de validación exógena para la detección de transbordos, utilizando datos de las encuestas London Travel Demand Surveys (LTDS) realizadas sobre una base de 12000 pasajeros en el año 2006. Los resultados deducidos a través de dicha validación manifestaron (*i*) la sobreestimación del número de viajes realizados por pasajero con respecto al derivado de las encuestas LTDS y (*ii*) la existencia de una cierta varianza entre el número de transbordos estimado por el método y el obtenido a partir de las encuestas. Utilizando diferentes valores para el parámetro MTT se determinaron desde 2.23 hasta 2.33 viajes diarios por pasajero frente a los 2.05 viajes establecidos a partir de las encuestas. Asimismo, se observaron diferencias del 6 %, -20 %, 7 % y 3 % entre los resultados de la estimación y los obtenidos de las encuestas para 1, 2, 3 y 4 transbordos diarios por pasajero, respectivamente.

Nassir et al. 2011 [41]

En este estudio se trabaja en la estimación de las paradas de bajada y la detección de transbordos de pasajeros en la red de autobuses de Minneapolis-Saint Pauls, Estados Unidos. Los datos disponibles se corresponden con transacciones obtenidas de un sistema AFC de tipo *entry-only*, en el que se registra la ubicación de la parada de subida y el instante temporal en el que el pasajero en cuestión subió al autobús.

Este estudio presenta una particularidad en el cálculo de la zona de *buffer* utilizada para estimar la parada de bajada más cercana a la parada de subida del siguiente segmento de viaje. La propuesta sugerida consiste en la multiplicación de la distancia euclídea existente entre las 2 paradas por $\sqrt{2}$, con el objetivo de adaptarlo a la distancia andando (*walking distance*). El razonamiento con el que justifica el uso de este factor de aumento es la presumible subestimación de la distancia real que se realiza andando debido a la configuración del esquema de conectividad entre calles de una ciudad. De esta forma, el estudio pretende estimar una distancia más parecida a la que se necesita recorrer verdaderamente para que un pasajero pueda llegar desde la parada de bajada hasta la siguiente de subida. En cuanto a la suposición de simetría del método *trip chaining*, esta

propuesta también asume que la parada de bajada del último segmento de viaje del día coincide exactamente con la parada de subida del primer segmento de viaje del día.

En lo relativo a los transbordos entre autobuses, el método propuesto emplea un valor de 90 minutos para el parámetro MTT, que determina el tiempo máximo para considerar que ha hecho un transbordo entre autobuses y no una actividad a mayores. Este estudio asume un valor de MTT mayor que la media de los demás estudios propuestos, con el objetivo de considerar como transbordo el segmento de la siguiente transacción acometida por la tarjeta de transporte, aunque el pasajero pueda haber tenido que realizar una actividad menor (ir al baño, comprar el periódico...) entre la parada de bajada de una transacción y la parada de subida de la siguiente. En caso de superarse el umbral establecido se deduce que efectivamente no se ha producido un transbordo sino una actividad mayor, por lo que no tendrá sentido conectar los segmentos de viaje asociados a las 2 transacciones correspondientes a la tarjeta del pasajero.

Por último, se lleva a cabo la validación del método utilizando datos AFC integrados con datos AVL y APC correspondientes a transacciones de viaje realizadas en la red de transporte de Minneapolis-Saint Pauls. Los resultados obtenidos muestran una precisión en la reconstrucción de los viajes origen-destino (correcta estimación de las paradas de subida y bajada) de alrededor del 60 % para los más de 80000 registros de transacciones considerados.

Wang et al. 2011 [29]

Esta propuesta trata la estimación de las paradas de autobús de bajada de los pasajeros en sus segmentos de viaje dentro de la red multimodal de transporte público de Londres, Reino Unido, compuesta por rutas de autobús y líneas de metro.

En este caso, las transacciones que se tienen de los viajes de metro son completas, incluyendo la información de subida y bajada de las estaciones, al provenir de un sistema AFC de tipo *entry-exit*. En cambio, el sistema AFC involucrado en las líneas de autobús es de tipo *entry-only*, pues solo se dispone de las transacciones asociadas a la subida de pasajeros en las que, además, está presente únicamente el instante de tiempo de subida y no la ubicación. Para paliar este inconveniente, en el estudio se lleva a cabo la integración del conjunto de datos de transacciones con los datos iBus de localización de autobuses, recogidos mediante un sistema AVL, para inferir la ubicación geográfica de las paradas de subida asociadas a cada transacción de viaje registrada.

En cuanto a los parámetros del método *trip chaining*, en este estudio se opta por utilizar una distancia umbral máxima de 1000 metros sobre las suposiciones de continuidad y simetría a la hora de estimar la parada de autobús de bajada de cada segmento de viaje a partir de la de subida del subsiguiente segmento.

En términos de la validación de la estimación realizada, se utilizan encuestas a pasajeros para construir sus flujos de origen y destino, de acuerdo con las rutas de autobús operativas. Debido a que se trata de un conjunto de datos con el que no se ha desarrollado el modelo de estimación, diremos que se ha realizado una validación exógena. Finalmente, los resultados obtenidos sobre el conjunto de datos de validación exponen una precisión del método en la estimación de las paradas de bajada del 66 % y 65 % en cada sentido de las rutas de autobús. Por otro lado, se refleja una diferencia del 7-8 % entre las paradas de autobús inferidas a través de la integración de los conjuntos AFC y AVL y las correspondientes a las encuestas de validación a los pasajeros.

Munizaga and Palma 2012 [30]

Este estudio trata la construcción de una matriz OD para la red multimodal de transporte público de Santiago, Chile, a partir de unas 70 millones de transacciones de subida en el transporte público recopiladas mediante un sistema AFC de tipo *entry-only* durante 1 semana de marzo de 2009 y otra de junio de 2010.

Para la inferencia de la información de subida a los autobuses, cuando las transacciones se registran dentro del propio autobús, este estudio, siguiendo el ejemplo de otras propuestas, lleva a cabo la integración del conjunto de transacciones AFC con datos AVL. Además, establece una distancia con un margen de 110 metros para atenuar la posible falta de precisión en la ubicación determinada por el sistema AVL. Adicionalmente, existe otra posibilidad a la hora de iniciar un viaje en autobús, que consiste en registrar la transacción en la parada de autobús, no dentro del mismo autobús. En este caso, la localización de subida puede ser inferida directamente sin necesidad de acometer la integración con datos AVL, propenso a fallos de precisión en la determinación de ma ubicación geográfica de los autobuses.

Este estudio propone un cambio en el método *trip chaining*, relativo a la suposición de simetría del método original. En este caso se considera un día virtual con diferente rango de horas que las habituales (de 00:00 a 23:59) y se propone considerar el inicio del día a las 4:00 y su correspondiente finalización a las 3:59 del siguiente. Así, el estudio pretende reducir el número de transacciones únicas asociadas a tarjetas de transporte con el objetivo de poder aplicar el método *trip chaining* a una mayor proporción de las transacciones, puesto que dicho método requiere un mínimo de 2 transacciones de pasajero por día para su uso.

Por otra parte, el estudio propone otra mejora del método orientada a la estimación adecuada de la parada de bajada de los pasajeros en las rutas de autobús de doble sentido teniendo en cuenta la subsiguiente parada de subida. La propuesta de esta mejora surge ante potenciales situaciones en las que la parada de bajada más cercana, en términos de distancia, a la parada de subida del siguiente segmento de viaje no es rentable desde el punto de vista del tiempo de viaje necesario para llegar a ella. Esta situación comentada

aparece descrita visualmente en la Figura 2.13. En ella se muestra un ejemplo de ruta en la que resulta más ventajoso bajarse en una parada intermedia de la ruta para llegar a la parada de subida siguiente que hacer todo el recorrido completo por esa ruta hasta ella. Para solventar esta situación, introduce el uso de una nueva función de coste a minimizar a la hora de estimar la parada de bajada de los autobuses de línea. La medida utilizada, en lugar del uso de un umbral de distancia máxima entre paradas, es el tiempo generalizado (TG), que se emplea para estimar la estación de bajada de un segmento de viaje que minimice su valor. Con esta nueva medida se pretende manejar el problema asociado a la desutilidad por parte del pasajero de viajar un mayor tiempo en autobús (*in-vehicle-travel-time*) que el que necesitaría si fuera andando (*walking time*) para llegar a una determinada parada.

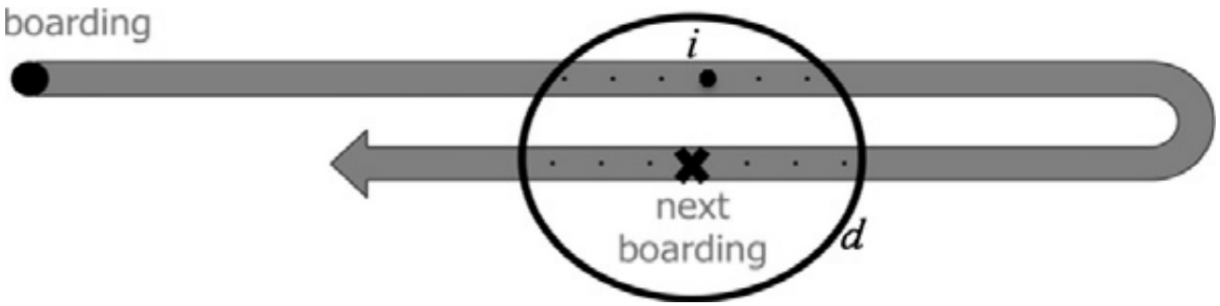

Figura 2.13: Situación en rutas de doble sentido referida en *Munizaga and Palma 2012*

La idea intuitiva detrás de este concepto puede ser explicada a través de su fórmula de cálculo:

$$Tg_i = t_i + f_w \cdot \frac{d_{i-post}}{s_w}$$

A partir de esta fórmula podemos expresar que el tiempo generalizado asociado a una parada de bus se define como el tiempo asociado a la parada (t_i) más un factor de penalización o desutilidad de la distancia andada necesario entre las paradas (f_w), obtenido a partir de una relación entre el tiempo andando necesario entre paradas (*walking time*) y el tiempo de recorrido necesario entre paradas por parte el autobús (*in-vehicle-travel-time*), que multiplica al resultado de dividir la distancia andando (*walking distance*) entre la parada de bajada a estimar y la parada de subida del siguiente segmento de viaje, por la velocidad media de andar (*average walking speed*). Finalmente, la parada que minimiza esta medida de tiempo generalizado es la estimada por el método como parada de bajada de ese segmento de viaje en bus. Cabe destacar que esta medida es utilizada solamente para estimar las paradas de bajada de los segmentos de viajes en autobús.

Por otro lado, a la hora de determinar el camino seguido por un pasajero dentro del transporte público, este estudio propone aplicar el algoritmo de Dijkstra, para hallar el camino más corto entre la parada de origen y la parada de destino de un viaje en la red de transporte público de Santiago.

En cuanto a la distancia umbral máxima considerada por este estudio a la hora de determinar las paradas de bajada candidatas de cada segmento de viaje en los demás modos de transporte, se ha optado por usar una distancia de 1000 metros de radio desde la siguiente parada de subida, la cual se interpreta como la distancia máxima que un pasajero está dispuesto a recorrer andando para llegar a su siguiente parada de subida.

En lo relativo a la detección de transbordos entre vehículos de transporte, este método emplea 18 minutos y 1000 metros como tiempo y distancia máximos entre paradas para considerar que se ha produccido un transbordo y no de una actividad.

Finalmente, el método propuesto ha sido validado de forma exógena mediante otro estudio posterior, *Munizaga et al. 2014*, en el que se utilizan datos de una matriz OD obtenida a partir de encuestas a pasajeros en el metro y datos de transacciones correspondientes a 53 pasajeros voluntarios. Los resultados derivados de la validación del método propuesto se anuncian como muy positivos, obteniéndose una precisión del 90 % en la estimación de los flujos origen-destino de tránsito de los pasajeros de muestra con los que se ha validado.

Gordon et al. 2013 [31]

Este estudio trata la estimación de las paradas de bajada y la detección de los transbordos producidos en la red de transporte público multimodal de Londres, Reino Unido. Los datos con los que trabaja se corresponden con transacciones de metro recogidas con un sistema AFC de tipo *entry-exit*, por lo tanto, con información de las subidas y bajadas de pasajeros de las líneas de metro, y transacciones de subida a autobuses de línea en las que solo se registra el instante de tiempo asociado. De forma semejante a [29] y [28], que también trabajan con transacciones de la red de transporte de Londres, se infieren las paradas de subida a los autobuses a través de la integración de las transacciones AFC con datos AVL de geoposicionamiento.

La estimación de las paradas de bajada se realiza con *trip chaining*, estableciendo una distancia umbral máxima de 1000 metros para la determinación de la zona de *buffer* alrededor de la parada de subida del siguiente segmento de viaje que contendrá las paradas de bajada candidatas entre las que hacer la estimación.

En cuanto a la detección de transbordos, este estudio asume un valor de 18 minutos para el parámetro MTT correspondiente con el tiempo requerido para caminar durante 750 metros a una velocidad media de 3 km/h, más un margen de 3 minutos adicionales, con el objetivo de acomodarse a diferentes tipos de accesos a las paradas/estaciones o velocidades de paso de las personas. Igualmente, de cara a inferir la ocurrencia de un transbordo o una actividad, el estudio considera que el pasajero lleva a cabo una actividad cuando deja pasar, sin subirse, el primer vehículo disponible de la ruta correspondiente a la seguida en el siguiente segmento de viaje. No obstante, esta consideración puede inducir a la subestimación de los transbordos verdaderos que se producen, pues en ciertos

momentos del tráfico puede denegársele a un pasajero el acceso a un vehículo de transporte por sobrepasar su ocupación máxima.

Por otro lado, se propone la utilización de diferentes distancias máximas de transbordo según el tipo de cambio entre modos de transporte dentro de la red multimodal, al considerar que son distintas las distancias medias a recorrer por un pasajero para moverse de un medio de transporte a otro. Comúnmente no se necesitará recorrer la misma distancia para moverse entre 2 líneas de autobús que para moverse del metro a una línea de autobús.

Así, la configuración propuesta por este estudio para la detección de transbordos entre segmentos de viaje se estructura en torno a 3 tests diferenciados: binario, temporal y espacial. En primer lugar, el test binario separa las transacciones que son las últimas del día, las que no disponen de una siguiente parada de subida correspondiente y las que no tienen una parada de bajada asignada al no haberse podido estimar. Por otro lado, el test temporal tiene en cuenta restricciones temporales como el uso de un parámetro MTT, dependiente de la distancia euclídea, que separa la parada de bajada y la subsiguiente parada de subida. Por último, el test espacial emplea el parámetro MTD, que representa la distancia máxima permitida entre paradas para considerar un transbordo, para evaluar si la distancia existente entre la parada de bajada y la subsiguiente parada de subida puede indicar el suceso de un transbordo. Una vez la transacción analizada haya pasado por los 3 tests, se considerará el segmento de viaje que describe como correspondiente a una actividad (*unlinked trip*) o como un transbordo asociado al anterior segmento de viaje (*linked trip*).

Por último, se llevó a cabo una validación exógena mediante encuestas a pasajeros para comprobar la precisión del método propuesto en la estimación de las paradas de bajada y en la detección de los transbordos. El estudio reporta una precisión del 74,5 % en la estimación de los tiempos y ubicaciones de bajada de los viajes en autobús. Por otro lado, entre las conclusiones más relevantes obtenidas de la validación se encuentra la sobreestimación en el número de viajes con 2 transbordos que se produjo con el método de detección de transbordos adoptado. Este hecho denota la necesidad de revisión de los parámetros del método para ajustarse mejor a la realidad.

Jun and Dongyuan 2013 [42]

Este estudio muestra la peculiaridad, con respecto al resto de estudios revisados, de solo considerar las transacciones de transporte registradas durante los picos de tráfico de la mañana y de la tarde, en los 5 días laborales de la semana. En este caso, los datos sobre los que trabaja se corresponden con transacciones realizadas dentro de la red de autobuses de la ciudad de Nanning, China.

El objetivo del estudio es determinar, a través del uso de patrones de tránsito, las matrices Origen-Destino asociadas a las horas punta de tránsito correspondientes a los movimientos laborales de entrada y salida del trabajo de los pasajeros. El método propuesto

se basa en la identificación de los pasajeros que viajan durante las horas punta de la mañana y de la tarde en los días laborables a través de una serie de suposiciones. En concreto, este método asume 3 suposiciones principales de partida:

- **Suposición 1**: Los pasajeros que regularmente toman el transporte público, tanto en las horas punta de la mañana como en las de la tarde de los días laborales, son los viajeros considerados.
- **Suposición 2**: Las paradas de subida comunes durante las horas pico de la mañana son consideradas las localizaciones de residencia de los viajeros.
- **Suposición 3**: Las paradas de subida comunes durante las horas pico de la tarde son consideradas las localizaciones de los lugares de trabajo de los viajeros.

Finalmente, las matrices OD resultantes se ofrecen como una buena información de entrada al análisis de la congestión del tráfico en momentos concretos.

Los resultados de evaluación del método propuesto en este estudio afirman que se logra una precisión en la estimación de las paradas de bajada del 83 %.

Nassir et al. 2015 [22]

Este estudio, plantea como objetivos principales la construcción de una matriz OD para conocer la frecuencia de viajes entre cada par de paradas de la red de transporte de South East Queensland, Australia, y la detección de transbordos, permitiendo la ocurrencia de actividades menores que no modifiquen el hecho de que se pueda considerar la realización de un transbordo entre paradas de bajada y subida consecutivas.

La particularidad más relevante de este estudio radica en el planteamiento de un algoritmo en 2 fases para la detección de transbordos. En la primera fase, las restricciones de espacio (MTD = 400 metros) y tiempo (MTT = 60 minutos) y de la misma ruta seguida en ambos segmentos de viaje son aplicadas para determinar la producción de una actividad o un transbordo. Mientras, la segunda fase emplea una serie de 5 criterios adicionales de tipo espacial y temporal (*gap*, *gap ratio*, *off-optimality*, *off-optimality ratio* y *circuity*) que demuestran ser capaces de determinar la realización por parte de los pasajeros de actividades menores entre las paradas. El estudio asigna valores a estos criterios de evaluación de acuerdo con los datos reales analizados de las transacciones de las tarjetas de la red de transporte de South East Queensland, al ser registradas con un sistema AFC de tipo *entry-exit*.

Por último, para la etapa de validación se emplearon tanto los datos de las transacciones del sistema AFC *entry-exit* como los asociados a la encuesta de viajes (*Household Travel Survey*) realizada a pasajeros durante el año 2009 para la red de transporte público de

South East Queensland. Así, el estudio declara una precisión del 98 % en la detección de actividades cortas frente a transbordos por parte del algoritmo en 2 fases propuesto.

Nunes et al. 2015 [3]

Este estudio se centra en la estimación de las paradas de bajada de los segmentos de viaje asociados a transacciones de tarjetas de pasajeros recopiladas a través de un sistema AFC de tipo *entry-only* instalado en las estaciones y paradas que constituyen la red de metro y autobuses de la ciudad de Oporto, Portugal.

De manera similar a [40] y [30], este estudio también propone la adaptación de la suposición de simetría del método *trip chaining* para considerar días virtuales, estableciendo las 5:00 como hora de inicio y las 4:59 del día siguiente como hora de finalización. Así, pretende adecuarse mejor a los hábitos de viaje de los pasajeros de Oporto y reducir el número de transacciones únicas que se observaron al considerar los rangos de hora normales. Dentro la configuración del mismo método, se consideró el valor de 640 metros para el parámetro de distancia umbral máxima desde la parada de subida del siguiente segmento de viaje, a partir del cual estimar la parada de bajada del segmento actual.

Otro aspecto destacado de este estudio surge en la limpieza de los datos de transacciones, donde considera como duplicadas y, por lo tanto, elimina las posibles segundas transacciones consecutivas que se registren dentro de un corto intervalo de tiempo y sigan la misma dirección de ruta del anterior segmento de viaje. Otros estudios, en cambio, optan por considerar dichas transacciones como correspondientes a un nuevo segmento de viaje independiente del anterior.

Finalmente, la evaluación del método adaptado de estimación de las paradas de bajada de los pasajeros ha sido realizada sobre un conjunto completo de transacciones registradas en la red de transporte en abril de 2010 y ha proporcionado un resultado de precisión del 62 %.

Alsger et al. 2016 [23]

Este estudio utiliza datos de transacciones de la red multimodal de transporte público de South East Queensland, Australia. En este caso, se opta por la utilización de una adaptación a la red de transporte considerada del método basado en reglas *trip chaining*, a través del cual se estiman las paradas de bajada de los pasajeros en los segmentos de viaje que recorren. En lo relativo a esta adaptación del método *trip chaining* se cuenta el establecimiento del valor de 530 metros para la distancia umbral máxima desde la parada de subida del siguiente segmento de viaje dentro de la cual buscar la parada de bajada a estimar para el segmento de viaje actual. La decisión de establecer dicho parámetro al

valor de 530 metros se justificó en el estudio al demostrar que se obtenía una mejora en la precisión de la estimación realizada por el método para la red de transporte público de South East Queensland (de un 66 % a un 72 %).

En cuanto a la detección de transbordos, gracias a la posibilidad de disponer de un conjunto de transacciones de tipo *entry-exit* para la validación del método, se llevó a cabo un análisis de sensibilidad con diferentes valores para el parámetro MTD (400, 800, 1000 y 1100 metros) con el que detectar los transbordos entre vehículos de transporte teniendo en cuenta sólo la información de subida al transporte público. De esta forma, a través de una validación endógena del método utilizando el conjunto de las transacciones de tipo *entry-exit* recolectadas, el estudio reportó la mejor precisión en el número de viajes estimado, en comparación con el real, en un 86 % de acierto para una distancia máxima de transbordo de 800 metros y un tiempo máximo de transbordo fijado en 60 minutos.

Jung and Sohn 2017 [43]

Este estudio analizado presenta la aplicación de un enfoque basado en *deep learning* para la estimación de las paradas de bajada de los segmentos de viaje que se realizan en la red multimodal de transporte público, compuesta por rutas de autobús y líneas de metro, de la ciudad de Seúl, Corea del Sur. En este análisis solo se utilizan las transacciones correspondientes a los viajes en autobús, de los cuales se conocen tanto la información de subida como la de bajada al utilizarse un sistema AFC de tipo *entry-exit* para el registro de la actividad de tránsito en el transporte público.

El hecho de contar con datos de transacciones en las que están presentes tanto la información de las paradas de subida como de las de bajada, habilita la posibilidad del uso de un enfoque basado en aprendizaje automático supervisado para la estimación de las paradas de bajada de los segmentos de viaje. Este estudio propone como aproximación de *deep learning* el uso de una red neuronal con 2 capas de neuronas ocultas y las funciones *Rectified Linear Unit* (ReLU) y *Softmax* como función de auxiliares de activación. Asimismo, establece un valor de *Drop Rate* del 30 % para intentar evitar el sobreajuste de la red neuronal. En este caso, los datos de entrada que son suministrados a la red neuronal se corresponden con 27 variables relativas tanto a las transacciones como a las características urbanas de cada zona (*land use*).

Finalmente, la validación endógena del modelo de red neuronal entrenado para la estimación de las paradas de bajada de los pasajeros se realiza a través de una partición del 20 % reservada del conjunto de transacciones con el que se ha hecho el entrenamiento de la red neuronal. El resultado de la evaluación exhibe una precisión en la estimación correcta de las paradas de bajada del 60 %, y de hasta el 87,5 % en caso de tolerar el fallo de la primera parada estimada y considerar la segunda más probable.

Kumar et al. 2018 [32]

Este estudio trata el problema de construcción de la matriz OD, presentando como principal peculiaridad la necesidad de estimar según la situación la parada de subida o de bajada correspondiente a cada transacción registrada. Los conjuntos de datos disponibles se dividen en 2 variantes, según esté disponible la información de subida (*entry-only*) o la de bajada (*exit-only*) de las paradas de la red de transporte público de Minneapolis-Saint Pauls, Estados Unidos, compuesta por rutas de autobús y líneas de metro.

Para el caso de la estimación de las paradas de subida, en este método integran los datos de transacciones con conjuntos de datos AVL de localización y GTFS (*General Transit Feed Specification*), de planificación horaria del tránsito. Por otro lado, relativo a la estimación de la información de las paradas de bajada, el estudio propone como novedad, empleando *trip chaining*, utilizar la primera parada de subida del día siguiente para estimar la parada de bajada del segmento de viaje asociado a las transacciones únicas del día anterior. Al igual que otros estudios, que proponen la suposición de días virtuales, en este caso también se pretende dar una estimación coherente de la parada de bajada cuando solo hay registrada una transacción en el día de rango horario normal.

En cuanto a la detección de transbordos, el estudio asigna un valor de 90 minutos para el parámetro de tiempo máximo de transbordo (MTT), que regula la consideración de un transbordo o una actividad entre paradas consecutivas. De manera equivalente a [41], este estudio asume un valor de MTT mayor que la media de los demás estudios revisados con el fin de habilitar la realización de una actividad menor por parte del pasajero sin necesidad de considerar que ha llegado al destino de su viaje.

Por último, el método propuesto fue validado de forma exógena a través de datos de encuestas tomadas en los propios medios del transporte público de la red en cuestión. El resultado reportado por el estudio en cuanto a la precisión media obtenida en la estimación de las paradas de subida o bajada, según el caso, fue de un 85 %.

Yan et al. 2019 [33]

Este estudio tiene como objetivo la estimación de las paradas de bajada asociadas a segmentos de viaje realizados en la red de autobuses de la ciudad de Shenzhen, China, a partir del uso de un conjunto de transacciones recogidas mediante un sistema AFC de tipo *entry-only*, que solo mantienen información del instante de tiempo en el que se produce la subida en los autobuses. De forma similar a otros estudios analizados, el enfoque planteado para estimar las ubicaciones de las paradas de subida asociadas a las transacciones registradas consiste en integrar los conjuntos de datos de transacciones con datos de ubicación de los autobuses, recopilados a través de un sistema AVL instalado en cada uno de ellos.

Como primera particularidad del estudio destaca la consideración de un umbral sobre el número de transacciones máximo que puede hacer una tarjeta para no ser considerada como de un trabajador de la red de transporte y, por lo tanto, eliminada del análisis. Así, se evita distorsionar los patrones de viaje seguidos por el resto de los pasajeros comunes. La segunda particularidad del estudio es relativa a la estimación de las paradas de bajada de autobuses en cada segmento de viaje, pues propone el uso novedoso de un algoritmo de estimación en 2 fases. Su funcionamiento se basa en la utilización del método común *trip chaining*, parametrizado con el valor de 1000 metros para la distancia umbral máxima, para estimar la información de bajada y el uso disjunto de diferentes técnicas de apendizaje automático (árboles de decisión, ensembles, Naive Bayes, SVM, KNN) para la estimación de las paradas de bajada asociadas a las transacciones en las que no se haya podido dar una estimación con el método *trip chaining*.

En cuanto a la detección de transbordos entre autobuses, este estudio propone el uso de 60 minutos y 1000 metros para los parámetros de tiempo (MTT) y distancia (MTD) máximas a la hora de considerar la ocurrencia de un transbordo entre paradas consecutivas, respectivamente. Por otro lado, al igual que el estudio [31], éste también considera la ocurrencia de una actividad en lugar de un transbordo cuando el pasajero deja pasar, sin subirse, el primer vehículo disponible de la ruta correspondiente a la seguida en el siguiente segmento de viaje. Hay que tener en cuenta que esta consideración en situaciones de congestión de tráfico, donde los autobuses vayan completos, puede inducir a la subestimación de los transbordos verdaderos, pues el pasajero realmente no tendrá la voluntad de desechar su subida en el autobús implicado sino simplemente no podrá subir en él.

Finalmente, se evalúa el método propuesto por el estudio mediante validación exógena, utilizando datos de 2 semanas de transacciones registradas a través de un sistema AFC de tipo *entry-exit* de la red de autobuses de Pekín, China. En este caso, el estudio estima que la aplicación del modelo desarrollado para los datos de la red de autobuses de Pekín será de utilidad para estimar adecuadamente las paradas de autobús de bajada correspondientes en la ciudad de Shenzhen.

Los valores de precisión en la estimación resultantes son segmentados para 2 grupos de usuarios según su frecuencia de viajes, denominados regulares e irregulares, que son establecidos en el estudio a través de la aplicación sobre los pasajeros de un proceso de clustering con el algoritmo *k-means*. De esta forma, el estudio informa de una precisión en la estimación de las paradas de bajada de un 74,4 % y un 70,2 % para los grupos de usuario regular e irregular, respectivamente.

Huang et al. 2020 [34]

Este estudio pretende la estimación de las matrices OD de tránsito que describen los flujos de pasajeros en la red de autobuses Suzhou, China. Los datos disponibles para el

estudio son transacciones de un sistema AFC de tipo *entry-only*, que solo registran la información temporal sobre las paradas de subida de los pasajeros de la red de autobuses. Con el fin de inferir de las paradas de subida a los autobuses de los pasajeros, el método propuesto aplica el algoritmo de clustering espacial *DBSCAN* sobre los datos del conjunto de transacciones AFC y los datos de localización GPS derivados de un sistema AVL instalado en cada autobús. Una vez se tienen estimadas las paradas de subida de los pasajeros, el estudio procede con la aplicación del método *trip chaining* para la estimación de las paradas de los autobuses donde se bajan los pasajeros en cada segmento de viaje.

En cuanto a la detección de transbordos, este método utiliza un valor de 1500 metros para la distancia máxima de transbordo e incluye una particularidad a la hora de establecer un valor variable para el tiempo máximo de transbordo basándose en la noción de la frecuencia de tránsito.

Por último, se lleva a cabo una validación exógena, empleando un método de distribución y recogida de billetes en la subida y a la bajada de los pasajeros de los autobuses, a partir de la cual se obtiene una precisión en la estimación de las paradas de subida y bajada de los pasajeros del 72-85 % y 70-86 %, respectivamente.

Assemi et al. 2020 [6]

Este estudio, al igual que [23], plantea la estimación de las paradas de bajada de los pasajeros en la red de transporte multimodal de South East Queensland, Australia, compuesta por autobuses, metro y ferries. Los datos de transacciones en esta red de transporte fueron recolectados a través de un sistema AFC de tipo *entry-exit*, donde se requiere a los pasajeros registrar su actividad tanto de subida como de bajada en los diferentes medios de transporte de la red. El hecho de disponer de información completa sobre los viajes realizados por los pasajeros habilita la posibilidad de emplear técnicas de aprendizaje supervisado con las que entrenar un modelo de estimación a partir de las transacciones disponibles.

Una suposición novedosa del método es la consideración de una distancia y tiempo mínimos de viaje de 0,03 kilómetros y 0,18 minutos, la cual es utilizada en la etapa de limpieza de datos para comprobar la existencia de posibles transacciones erróneas debido a motivos humanos como la evasión de pago o pasar la tarjeta y finalmente no subirse a un vehículo por haberse equivocado.

La peculiaridad principal del método del estudio radica en el método empleado para estimación de las paradas de bajada de los pasajeros. Este estudio considera el uso de *deep learning*, como mecanismo para mejorar el rendimiento ofrecido por el método determinista *trip chaining*, que venía siendo propuesto por anteriores estudios de investigación. Como motivación para su uso, el estudio destaca el considerable aumento obtenido en la precisión del modelo de estimación con respecto al que se conseguiría utilizando *trip chaining*.

Siguiendo este enfoque, el conjunto de datos de transacciones puede ser dividido en 2 partes para llevar a cabo el desarrollo/entrenamiento y una validación endógena del modelo de estimación utilizando un mismo conjunto de datos. En cuanto al entrenamiento de la red neuronal utilizada, las variables que se incluyen como entrada son la ubicación e instante de tiempo de subida, el número de paradas y la distancia existentes entre la parada de subida y cada una de las potenciales paradas de bajada, un booleano para indicar si el segmento de viaje actual es el último del día y la duración estimada entre la parada de subida y cada potencial parada de bajada. En este caso, la variable de salida que predice la red neuronal será la parada de bajada más probable del segmento de viaje descrito por los datos de entrada suministrados a la red neuronal.

Finalmente, los resultados de precisión en la estimación de las paradas de bajada reportados por el estudio demuestran una cierta mejoría debido al uso de redes neuronales, alcanzándose una precisión de hasta el 79,5 %, frente al 72,2 % que se lograba con el método *trip chaining*.

Cheng et al. 2020 [7]

Este estudio trabaja sobre datos de transacciones registradas a partir de un sistema AFC de tipo *entry-exit* de la red de metro de Guangzhou, China, con el objetivo de estimar las estaciones de metro de bajada de cada uno de los segmentos de viaje realizados por los pasajeros.

Para la estimación de las paradas de bajada este estudio opta por el uso de un modelo de estimación probabilístico basado en asignación latente de Dirichlet (*latent Dirichlet allocation*) con el objetivo de relajar las restrictivas suposiciones que se toman en los métodos basados en reglas, como el método *trip chaining*. El modelo propuesto se entrena para la estimación de las paradas de bajada de los pasajeros con la información de ubicación y tiempo de las paradas de subida asociadas a transacciones de tarjetas de transporte recopiladas durante un periodo de 3 meses.

En la evaluación del modelo de estimación, este estudio reporta un incremento del 2 % en el porcentaje de estimaciones correctas en el análisis de las transacciones por pasajero con respecto al que se obtendría con el uso del método más tradicional *trip chaining*.

Para la evaluación del modelo de estimación, este estudio compara el rendimiento del método propuesto con otros modelos de referencia *benchmarks* basados en información histórica de los destinos de viaje de los pasajeros. El resultado de esta evaluación muestra un incremento medio del 2 % en la precisión de la estimación usando el método propuesto, en comparación con la precisión conseguida con los modelos de referencia utilizados.

Lee et al. 2022 [44]

Este estudio se centra en mejorar la estimación de las paradas de bajada asociadas a segmentos de viaje realizados en la red de autobuses de la ciudad de Sejong, Corea del Sur. Las transacciones que describen los segmentos de viaje recorridos por los pasajeros en la red de transporte público son recopiladas a través de un sistema AFC de tipo *entry-exit*, lo cual permite efectuar una validación endógena del método propuesto.

El propósito principal de este estudio es desarrollar un método que permita determinar las paradas de bajada de los segmentos de viaje que no hayan podido ser estimadas mediante el método *trip chaining*. Para ello, se necesita disponer de más información que permita caracterizar mejor la realidad de viajes de los pasajeros en la red de transporte público. Así, se propone el uso de patrones de viaje temporales junto con información histórica de registros de subida de los pasajeros en la red de transporte correspondientes al período comprendido entre el 1 de abril y el 31 de mayo de 2018. Además, se asumen las dos siguientes suposiciones básicas a la hora de establecer los patrones de viaje temporales, que son:

- **Suposición 1**: Las paradas de bajada de los segmentos de viaje estarán ubicadas cerca de paradas donde el pasajero sube al transporte público frecuentemente.
- **Suposición 2**: Las paradas de bajada de los segmentos de viaje dependen de la hora de subida.

La determinación de estos patrones de viaje temporales se estima a través de un modelo de mezcla de gaussianas (*gaussian mixture model*) aplicado sobre unas agrupaciones de pasajeros resultantes de un proceso previo de clustering, realizado mediante el método *k-means* a partir de los datos de transacciones de los pasajeros en el período considerado.

En cuanto a la validación del método propuesto, el estudio reporta un precisión del 74,9 % en la estimación de las paradas de bajada cuando se combina el método *trip chaining* con los patrones de viaje temporales, destinados a estimar las paradas de bajada de los segmentos de viaje que no pudieron ser asignadas mediante *trip chaining*. A este respecto, el estudio informa de una mejora del 14,9 % en la estimación al hacer la comparación con la precisión del 60 % obtenida con el método *trip chaining* original.

Tabla Comparativa de propuestas

A continuación, se presenta una tabla comparativa general que resume las principales características de cada una de las propuestas analizadas, incluyendo la correspondiente a este trabajo (véase la Tabla 2.9).

Tabla 2.9: Matriz Comparativa de Propuestas Analizadas

Propuesta	**Ubicación**	**Modos Transporte**			**Tipo AFC**	**Salida**	**Método**
		Bus	Metro / Tren	Ferry			
ACTUAL	Madrid, 2022	✓	✓	✗	Entry-only	Matriz OD multimodal	Trip chaining
Barry et al. 2002	Nueva York, Estados Unidos	✗	✓	✗	Entry-only	Matriz OD estación-estación	Trip chaining
Trépanier et al. 2007	Gatineau, Canadá	✓	✗	✗	Entry-only	Estimación de las paradas de bajada	Trip chaining
Zhao et al. 2007	Chicago, Estados Unidos	✓	✓	✗	Entry-only	Matriz OD estación-estación y detección de trans-bordos tren-tren y tren-autobús	Trip chaining
Farzin 2008	Sao Paulo, Brasil	✓	✗	✗	Entry-only	Matriz OD	Trip chaining
Barry et al. 2009	Nueva York, Estados Unidos	✓	✓	✗	Entry-only	Matriz OD multimodal basada en zonas	Trip chaining
Seaborn et al. 2009	Londres, Reino Unido	✓	✓	✗	Entry-only	Matriz OD multimodal y detección de trans-bordos intermoda-les	Trip chaining

Continúa en la página siguiente

Tabla 2.9 – continuación de la página anterior

Propuesta	**Ubicación**	**Modos Transporte**			**Tipo AFC**	**Salida**	**Método**
		Bus	Metro / Tren	Ferry			
Nassir et al. 2011	Chicago, Estados Unidos	✓	✗	✗	Entry-only	Estimación de las paradas de bajada y detección de trans-bordos	Trip chaining
Li et al. 2011	Jinan, China	✓	✗	✗	Entry-only	Estimación de las paradas de bajada	Trip chaining
Wang et al. 2011	Londres, Reino Unido	✓	✓	✗	Entry-only	Estimación de las paradas de bajada	Trip chaining
Munizaga and Palma 2012	Santiago, Chile	✓	✓	✗	Entry-only	Matriz OD multimodal	Trip chaining
Gordon et al. 2013	Londres, Reino Unido	✓	✓	✗	Entry-only	Inferencia de viajes in-termodales	Trip chaining y detección de transbordos basada en reglas
Jun and Dong-yuan 2013	Nanning, China	✓	✗	✗	Entry-only	Matriz OD	Método estadístico de 3 suposiciones
Nassir et al. 2015	South East Queens-land, Australia	✓	✓	✓	Entry-exit	Matriz OD multimodal y detección de trans-bordos con actividades menores	Trip chaining y método de detección de transbordos con actividades menores

Continúa en la página siguiente

Tabla 2.9 – continuación de la página anterior

Propuesta	**Ubicación**	**Modos Transporte**			**Tipo AFC**	**Salida**	**Método**
		Bus	Metro / Tren	Ferry			
Nunes et al. 2015	Oporto, Portugal	✓	✓	✗	Entry-only	Matriz OD multimodal	Trip chaining
Alsger et al. 2016	South East Queensland, Australia	✓	✓	✓	Entry-exit	Estimación de las paradas de bajada	Trip chaining
Jung and Sohn 2017	South East Queensland, Australia	✓	✓	✓	Entry-exit	Estimación de las paradas de bajada	Deep Learning
Kumar et al. 2018	Minneapolis-Saint Pauls, Estados Unidos	✓	✓	✗	Entry-only / Exit-only	Matriz OD multimodal	Trip chaining
Yan et al. 2019	Shenzhen, China	✓	✗	✗	Entry-only	Estimación de las paradas de bajada	Trip chaining y aprendizaje automático
Huang et al 2020	Suzhou, China	✓	✗	✗	Entry-only	Matriz OD	Trip chaining
Assemi et al. 2020	South East Queensland, Australia	✓	✓	✓	Entry-exit	Estimación de las paradas de bajada	Deep Learning
Chen et al. 2020	South East Queensland, Australia	✓	✓	✓	Entry-exit	Estimación de las paradas de bajada	Método Probabilístico (LDA)
Lee et al. 2022	Sejong, Corea de Sur	✓	✗	✗	Entry-exit	Estimación de las paradas de bajada	Trip chaining y patrones de viaje temporales

También, se han resumido en otra tabla comparativa (véase la Tabla 2.10) las diferentes estrategias y particularidades asumidas por las propuestas anteriores que utilizan *trip chaining* como método base de estimación. Partiendo de las suposiciones planteadas por el

método *trip chaining* original, estas ideas de adaptación de *trip chaining* propuestas por los estudios anteriores nos aportan una visión general de las distintas posibilidades que existen de cara a estimar las paradas de bajada de los pasajeros. Así, tendremos una guía de las diferentes opciones ya utilizadas en la literatura que nos puede ayudar a la hora de tomar decisiones con respecto a nuestra aplicación de *trip chaining* en el contexto de la red de transporte público de Madrid.

Tabla 2.10: Comparativa de Propuestas de Trip Chaining

Propuesta	**Adaptación de Suposiciones**		**Integración Datos**	**Consideraciones Adicionales**
	Continuidad	Simetría		
Barry et al. 2002	✗	✗	✗	Es el primer estudio en aplicar el método trip chaining.
Trépanier et al. 2007	✗	Distancia de cercanía entre paradas de fin e inicio en lugar de coincidencia exacta.	✗	Es el primer estudio en adaptar la suposición de simetría diaria.
Zhao et al. 2007	✗	✗	Integración con datos AVL	Uso de un valor de MTD igual a 400 metros para la detección de transbordos.
Farzin 2008	✗	✗	Integración con datos AVL	Uso de datos APC para estimar la matriz OD.
Barry et al. 2009	✗	Distancia de cercanía entre paradas de fin e inicio en lugar de coincidencia exacta; Consideración de día virtual (rango horario de 3:00 a 2:59)	Integración con datos horarios	Uso de un valor de MTT igual a 18 minutos para la detección de transbordos.

Continúa en la página siguiente

Tabla 2.10 – continuación de la página anterior

Propuesta	**Adaptación de Suposiciones**		**Integración Datos**	**Consideraciones Adicionales**
	Continuidad	Simetría		
Seaborn et al. 2009	✗	✗	Integración con datos AVL y de paradas	Diferentes rangos de valores de MTD y MTT según el cambio entre tipos de modos de transporte.
Nassir et al. 2011	Cálculo de *walking distance* multiplicando la distancia euclídea por $\sqrt{2}$	✗	✗	Uso de un valor de MTT igual a 90 minutos para la detección de transbordos.
Wang et al. 2011	✗	Distancia de cercanía entre paradas de fin e inicio en lugar de coincidencia exacta	Integración con datos AVL	Uso de un valor de *walking distance* igual a 1000 metros.
Munizaga and Palma 2012	Uso del Tiempo Generalizado (TG), en lugar de un umbral de *walking distance*, para los viajes en autobús	Distancia de cercanía entre paradas de fin e inicio en lugar de coincidencia exacta; Consideración de día virtual (rango horario de 4:00 a 3:59)	Integración con datos AVL	Uso de un valor de *walking distance* igual a 1000 metros; Uso de valores de MTD Y MTT iguales a 1000 metros y 18 minutos, respectivamente.
Gordon et al. 2013	✗	Distancia de cercanía entre paradas de fin e inicio en lugar de coincidencia exacta	Integración con datos AVL	Uso de un valor de *walking distance* igual a 1000 metros; Uso de valor de MTT igual a 18 minutos; Estructura de 3 tests para la detección de transbordos.

Continúa en la página siguiente

Tabla 2.10 – continuación de la página anterior

Propuesta	**Adaptación de Suposiciones**		**Integración Datos**	**Consideraciones Adicionales**
	Continuidad	Simetría		
Nunes et al. 2015	✗	Distancia cercanía entre paradas de fin e inicio en lugar de coincidencia exacta; Consideración de día virtual (rango horario de 5:00 a 4:59)	Integración con datos AVL	Uso de un valor de *walking distance* igual a 640 metros.
Alsger et al. 2016	✗	Distancia de cercanía entre paradas de fin e inicio en lugar de coincidencia exacta	✗	Uso de un valor de *walking distance* igual a 530 metros; Uso de valores de MTD Y MTT iguales a 800 metros y 60 minutos, respectivamente.
Kumar et al. 2018	✗	Consideración de la primera parada de subida del día siguiente para estimar la última de parada de bajada del día	Integración con datos AVL y GTFS	Uso de un valor de MTT igual a 90 minutos para la detección de transbordos.
Yan et al. 2019	✗	Distancia de cercanía entre paradas de fin e inicio en lugar de coincidencia exacta; Consideración alternativa de la primera parada de subida del día siguiente	Integración con datos AVL	Establecimiento de un número máximo de transacciones permitidas en un día; Uso de un valor de *walking distance* igual a 1000 metros.; Uso de valores de MTD Y MTT iguales a 1000 metros y 60 minutos, respectivamente; Uso alternativo a *trip chaining* de técnicas de aprendizaje automático.

Continúa en la página siguiente

Tabla 2.10 – continuación de la página anterior

Propuesta	**Adaptación de Suposiciones**		**Integración Datos**	**Consideraciones Adicionales**
	Continuidad	Simetría		
Huang et al. 2020	✗	Distancia de cercanía entre paradas de fin e inicio en lugar de coincidencia exacta	Integración con datos AVL	Uso de valor de MTD igual a 1500 metros.
Lee et al. 2022	✗	Uso de datos históricos de primera subida	✗	Uso adicional a *trip chaining* de patrones de viaje temporales.

Conclusiones del Estado del Arte

Cada una de estas propuestas emplea un método para la resolución del problema adaptado a la topología de la red de transporte propia de la ciudad en la que se centran estos estudios. Por ello, cada estudio asume unas ciertas suposiciones, acordes con las peculiaridades del sistema sobre el trabajan, a la hora de desarrollar el modelo de estimación que se utilizará para reconstruir los viajes realizados por los pasajeros, a partir de los cuales se elabora la matriz OD asociada con la que realizar posteriores análisis para otros propósitos mayores. Estas peculiaridades, alineadas con la topología de la red de transporte, pueden afectar desde el simple cambio de valor de ciertos parámetros de entrada al modelo hasta la consideración de la adaptación del método para tener en cuenta el distinto flujo de operación entre los diferentes medios de transporte que constituyen una red multimodal de transporte de pasajeros, los cuales pueden comprender diferentes porciones de información a la hora de registrar las transacciones de viaje de un pasajero en los mismos.

Como se ha comprobado, el enfoque de trabajo de las propuestas presentadas es variopinto, optando sus autores por el uso de métodos estadísticos, variantes mejoradas de los mismos, métodos probabilísticos y hasta el uso de técnicas de *deep learning* con el objetivo de aprovechar las capacidades del aprendizaje automático (*machine learning*) para mejorar la eficacia en la estimación del modelo desarrollado. No obstante, debido a que frecuentemente no se dispone de un número suficiente de datos correctamente etiquetados en este ámbito, se ha observado que la utilización de técnicas de aprendizaje supervisado no es aún algo habitual.

Entre las propuestas estudiadas también se ha comprobado como varía el tipo de información que debe ser estimada antes de poder emprender la construcción de las

matrices de tráfico OD. Este hecho estará condicionado por la información recopilada en las transacciones registradas por los terminales AFC de los medios de transporte de la red considerada. Dependiendo de si la información desconocida es la correspondiente a la parada de subida o a la de bajada, diferentes aproximaciones son empleadas para solventar este problema y poder estimar su valor más acertado.

Para la resolución de este problema de estimación, tiende a ser necesario el uso e integración de diversas fuentes de datos adicionales que complementen a los propios registros de transacciones recogidos por los sistemas de automáticos de tarificación (AFC). La utilización de estas otras fuentes de datos en las propuestas analizadas dependerá de las consideraciones tenidas en cuenta en las mismas para implementar el método de estimación que proponen. Generalmente, las principales fuentes de datos utilizadas en la literatura para su integración con los conjuntos de transacciones provenientes de los sistemas AFC con el objetivo de poder abordar la estimación de la información desconocida que se necesite averiguar son: AVL, GTFS, APC o datos de horarios de tránsito.

Por otra parte, en los diferentes estudios se ha expuesto la necesidad de realizar procesos de limpieza de datos (*data cleaning*) con los que eliminar posibles errores, información duplicada o resolver inconsistencias presentes en los datos recopilados de las fuentes de datos disponibles. Asimismo, también se pone de manifiesto la importancia de contar con grandes conjuntos de datos adicionales con los que llevar a cabo una validación efectiva del modelo propuesto, lo cual no siempre se encuentra a disposición de los investigadores. En cuanto a la validación de los pasos del método de estimación propuesto en cada uno de los estudios, se ha podido observar como la mayoría de ellos emplean validación exógena o externa, basada en conjuntos de datos externos sobre los que no se ha desarrollado el método, como encuestas realizadas por agencias de transporte público, frente a la validación endógena o interna, consistente en utilizar parte del mismo conjunto de datos utilizado en el desarrollo del modelo de estimación. Generalmente, este hecho se debe a la dificultad de disponer de conjuntos de transacciones de tipo *entry-exit* que, además, sean lo suficientemente grandes como para proporcionar una validación robusta de las técnicas utilizadas.

Finalmente, en la mayoría de los casos no se puede concluir que unos sean mejores que otros, pues la complejidad de la red de transporte sobre la que trabajan y la cantidad/calidad de los datos de entrada de los que disponen son diferentes en cada situación.

Propuesta Final Planteada

Una vez revisadas las propuestas existentes de otros autores y teniendo en cuenta la topología de la red de transporte público de Madrid y los datos de los que se disponen para realizar nuestro estudio, es decir, transacciones de los diferentes modos de transporte de la red recogidas mediante un sistema AFC de tipo *entry-only*, se decide optar por la implementación del método *trip chaining*, adaptando sus suposiciones y parámetros a las

peculiaridades de la red de transporte público de Madrid sobre la que se va a aplicar, para la estimación de las paradas de bajada de los pasajeros y la posterior reconstrucción de los flujos de tránsito origen-destino recorridos por los mismos a lo largo de la red.

En lo relativo a los transbordos entre vehículos de un mismo o distintos medios de transporte dentro de la red, se considerarán diferentes estrategias basadas en reglas adaptadas también a las condiciones de la red de transporte de Madrid. Igualmente, los valores de los parámetros correspondientes a la tarea de detección de transbordos serán seleccionados teniendo en cuenta las consideraciones expresadas en estudios previos acordes al actual que proponemos.

Planificación y Presupuesto

En este capítulo se describirán, a grandes rasgos, tanto la metodología de trabajo utilizada para llevar a cabo la organización del trabajo en el proyecto como la metodología de desarrollo adoptada para seguir un ciclo de vida de proyecto *Big Data*. A continuación, se hará una estimación del esfuerzo necesario para acometer las tareas que satisfarán los objetivos planteados en el alcance del proyecto. Seguidamente, se especificará la planificación temporal del proyecto dentro de los límites temporales impuestos por las fechas establecidas para la realización del Trabajo de Fin de Máster. Por último, se presentará el presupuesto económico dispuesto para el proyecto, el cual estará condicionado por la planificación temporal indicada previamente, además de por otros recursos necesarios para su desarrollo.

3.1. Metodología de Trabajo

La metodología que se ha utilizado para organizar el trabajo a realizar durante el proyecto ha sido **UVagile** [45]. Esta es una metodología creada dentro de la Universidad de Valladolid que, basándose en los principios del marco de trabajo Scrum [46], aborda la organización ágil de procesos de enseñanza-aprendizaje con el objetivo de fomentar el aprendizaje incremental del alumno y la retroalimentación frecuente por parte del profesor.

Dentro del área de los trabajos de fin de estudios en el ámbito universitario, tales como los Trabajos de Fin de Grado y Máster, UVagile propone adoptar una forma de trabajo que se asemeje a la seguida en proyectos profesionales con equipos ágiles, es decir, estableciendo una dinámica de trabajo iterativa e incremental que sea mantenida a lo largo de todo el proyecto. Para ello, UVagile define una serie de roles, eventos y artefactos, inspirados en Scrum, que lo caracterizan y que deben ser explicados para favorecer su entendimiento.

Roles

En primer lugar, se describirán los roles involucrados cuya participación se plantea dentro de la dinámica de trabajo de UVagile.

- **Estudiante**: representa el rol principal dentro de UVagile, siendo desempeñado por el alumno y asemejándose al rol Developer de Scrum. Sus principales responsabilidades son la construcción del producto final a entregar y su defensa ante el tribunal.
- **Profesor**: representa el rol desempeñado por el tutor del proyecto, cuya misión es servir de guía al estudiante para asegurar que alcanza los objetivos de aprendizaje marcados y maximizar la calidad del producto final que entrega. Este rol está inspirado en los de Product Owner y Scrum Master de Scrum.
- **Comunidad**: representa el conjunto de los estudiantes que realizan sus trabajos de fin de estudios de manera simultánea y que contribuyen a establecer el entorno de aprendizaje colectivo en el que cada estudiante desarrolla su proyecto. Su misión es conformar un grupo de personas en el que la retroalimentación mutua beneficie a todos los estudiantes en sus proyectos individuales con el fin de favorecer su proceso de desarrollo.
- **Tribunal**: es la comisión evaluadora del trabajo de fin de estudios ante la cual el estudiante debe realizar su defensa. Sus responsabilidades vendrán definidas por el reglamento específico que rija en cada caso. El equivalente de este rol en Scrum sería la figura del cliente.

Eventos

Los eventos representan el punto de conexión y comunicación entre los diferentes roles definidos en la metodología de trabajo UVagile. Su misión es establecer un proceso de trabajo continuo y transparente que favorezca el seguimiento de los progresos producidos por el estudiante. Estos eventos son períodos de tiempo limitado (*time-boxes*) con los que se busca afianzar una regularidad en el proceso de desarrollo evitando así la necesidad de programar reuniones no definidas. Dentro de ellos se llevan a cabo todas las tareas involucradas en el proyecto.

- **Sprint**: es el evento principal que actúa como contenedor del resto. Su función es la de englobar todo el trabajo necesario para construir un incremento de producto. Un proyecto se compone de varios sprints de igual duración en los que se producen siempre todos los eventos, favoreciendo la predictibilidad del proceso y creando una rutina de trabajo. Para este proyecto se han programado 4 sprints de 3 semanas de duración cada uno.
- **Reunión de Inicio**: es equivalente al evento *Sprint Planning* de Scrum. Este es el evento con el que da comienzo cada sprint del proyecto y en el que participan tanto el estudiante como el profesor. Su misión es consolidar el objetivo del sprint y establecer la planificación del trabajo a realizar en el mismo. En cada Reunión de Inicio se deciden las nuevas historias de proyecto sobre las que se trabajarán a lo largo del sprint, así como las mejoras derivadas de las correcciones del incremento

	Lunes	**Martes**	**Miércoles**	**Jueves**	**Viernes**
Semana 1		Reunión de Inicio			
Semana 2		Reunión de Sincronización			
Semana 3		Reunión de Sincronización			
Semana 4		Comunicación de Progresos			

Tabla 3.11: Esquema temporal de un Sprint en UVagile

presentado al finalizar el sprint inmediatamente anterior. Además, el estudiante tiene la responsabilidad de planificar las tareas que abordará para satisfacer cada una de las historias de proyecto incluidas en el alcance del sprint.

- **Reunión de Sincronización**: está inspirada en el evento *Daily Scrum* de Scrum. Este evento representa una reunión semanal de corta duración (máximo de 15 minutos) en la que el estudiante expone al profesor el trabajo realizado desde la última reunión, comunicando los avances y/o bloqueos que han podido surgir. Asimismo, se expone el trabajo a realizar en adelante y se comentan posibles soluciones a los impedimentos surgidos. En este proyecto las reuniones de sincronización tenían lugar todos los martes por la mañana con el fin de seguir una rutina de trabajo.

- **Comunicación de Progresos**: está inspirada en el evento *Scrum Review* de Scrum. La comunicación de progresos constituye la oportunidad para el estudiante de presentar el incremento de producto construido hasta el momento y preparar el acto de defensa. Esta reunión tiene lugar al final de cada Sprint y participan en ella el estudiante, el profesor y la comunidad, con el fin de agrupar una masa crítica que aporte una retroalimentación valiosa que posibilite la corrección y mejora del siguiente incremento a construir.

- **Retrospectiva**: es equivalente al evento *Scrum Retrospective* de Scrum. Este es el último evento del Sprint y consiste en la evaluación del proceso de trabajo seguido durante el Sprint. En esta reunión participan el estudiante, el profesor y la comunidad con el objetivo de detectar los aspectos positivos y negativos del proceso de trabajo seguido durante el Sprint para potenciarlos o corregirlos de cara al siguiente Sprint según cada caso. La mecánica de esta reunión se basa en cuestionar qué se hizo bien, qué se hizo mal y qué posibles mejoras se pueden implementar para el siguiente Sprint.

Un esquema temporal de la planificación de los eventos de un Sprint según la metodología UVagile puede verse en la Tabla 3.11.

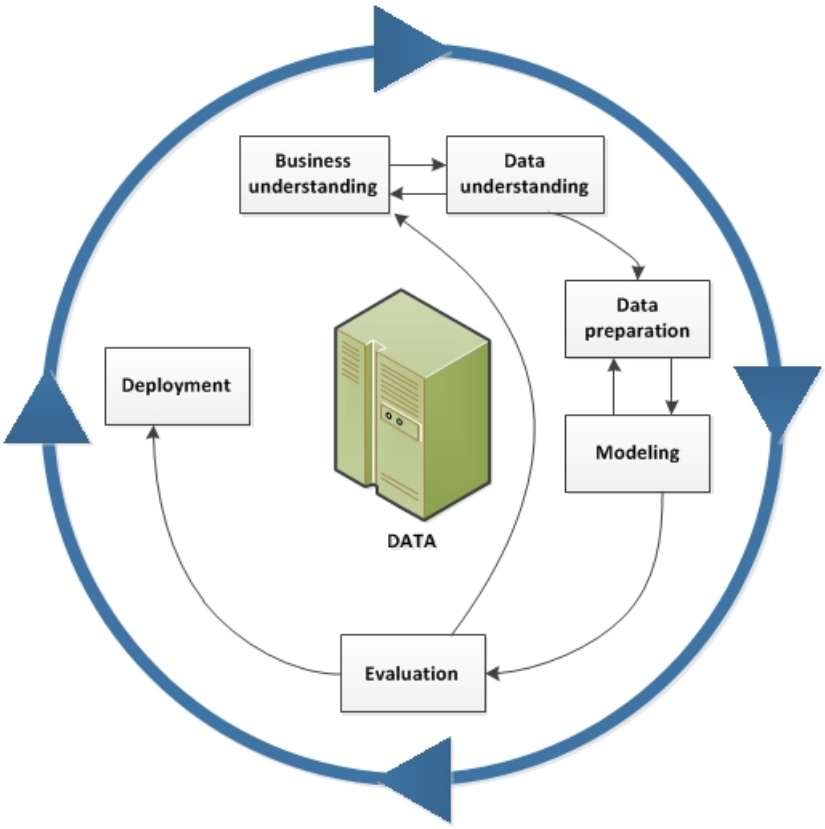

Figura 3.14: Ciclo de vida de un proyecto según CRISP-DM [12].

3.2. Metodología de Desarrollo

Para la realización de este proyecto se ha seguido una adaptación de la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) [47], una referencia habitual que rige las fases a seguir en el desarrollo de proyectos de ciencia de datos con el objetivo de completarlos de manera exitosa.

CRISP-DM es un marco de trabajo designado para proyectos de software que tratan con datos y cuyo objetivo es extraer valor a partir de ellos. Esta metodología resulta de interés porque define un conjunto de fases y tareas generales que se adecúan a las características de la mayor parte de los proyectos que se suelen abordar en el área de la minería de datos. Un resumen visual de las fases del proceso de desarrollo propuesto por CRISP-DM puede verse en la Figura 3.14. Esta figura representa un ciclo de vida constituido por 6 fases diferenciadas que abarcan el proceso completo de desarrollo de un proyecto de ciencia de datos. Dichas fases se pueden interrelacionar de forma bidireccional, habilitando la opción de iterar entre ellas para aprovechar en fases anteriores la retroalimentación proporcionada al realizar otras fases posteriores. Así, la principal ventaja de este ciclo de vida radica en la posibilidad de retornar a fases anteriores del proceso según sea necesario. A continuación, se aporta una descripción concisa de cada una de estas fases del ciclo de vida.

- **Comprensión del negocio (*Business understanding*)**: esta primera fase consiste en el entendimiento de las necesidades de negocio y su alineamiento con objetivos de proyecto, de manera que queden definidos los principales objetivos que se deben satisfacer al finalizar el mismo.
- **Comprensión de los datos (*Data understanding*)**: consiste en la exploración y entendimiento de los datos disponibles con los que trabajar para la consecución de los objetivos de proyecto planteados. Debe proporcionar una visión general de los datos y las posibilidades de extracción de valor que éstos pueden ofrecer.
- **Preparación de datos (*Data preparation*)**: consiste en la realización de las transformaciones necesarias sobre los datos previas a su modelado, de acuerdo con las necesidades propias del proyecto.
- **Modelado (*Modelling*)**: consiste en el análisis, selección y aplicación de las técnicas de modelado de datos más apropiadas de acuerdo con los objetivos del proyecto, con el fin de derivar los resultados a evaluar en la siguiente fase.
- **Evaluación (*Evaluation*)**: consiste en el análisis y revisión de los resultados finales para determinar la validez de los modelos de datos construidos en la fase anterior. En esta fase se decide si los resultados obtenidos satisfacen los objetivos de proyecto planteados o si, por el contrario, se debe retornar a fases previas del ciclo de vida para mejorar la propuesta de solución actual.
- **Despliegue (*Deployment*)**: esta última fase consiste en la explotación de los resultados finales del proyecto y su puesta a disposición para el usuario final.

Una vez comentadas las bases de esta conocida metodología, se plantea una adaptación de la misma para abordar la realización de este proyecto Big Data. Esta propuesta de adaptación es sintetizada en el ciclo de vida de la figura 3.15, cuyas fases del proceso se ajustan mejor a las necesidades comunes a este tipo de proyectos. Este ciclo de vida está compuesto por 5 fases diferentes que se pueden interrelacionar de forma bidireccional entre sí, posibilitando la iteración sobre las mismas, con el objetivo de aprovechar una retroalimentación entre fases que resulte beneficiosa para la realización efectiva del proyecto. A continuación, se procede a describir, en referencia a nuestro proyecto, cada una de estas fases del ciclo de vida.

- **Inicio**: esta primera fase comprende el establecimiento de los fundamentos técnicos, la definición de las necesidades de negocio y la formalización de los objetivos de proyecto a abordar. Asimismo, incluye las tareas de planificación asociadas al proyecto.
- **Análisis**: esta fase consiste en la exploración y estudio de los datos disponibles (*raw data*) y en el diseño del conjunto de transformaciones necesarias para convertir dichos datos iniciales en los datos finales requeridos por las necesidades de negocio (*smart*

Figura 3.15: Ciclo de vida de un proyecto Big Data. Adaptado de [13]

data). El resultado de esta fase será la consolidación de una primera abstracción del proyecto.

- **ETL**: esta fase implementa las tareas de extracción de datos iniciales, transformación de los mismos de acuerdo con el diseño anterior y carga de los datos transformados (*smart data*) para su evaluación posterior.
- **Evaluación**: esta fase trata la utilización del *smart data* para determinar el cumplimiento de los objetivos planteados en el proyecto.
- **Despliegue**: esta última fase consiste en la explotación de los resultados del proyecto de cara a su uso en producción por parte de los usuarios finales. Se debe tener en cuenta que este proceso será muy dependiente del entorno de explotación concreto en el que se requerirá desplegar los resultados finales obtenidos del proceso. Sin embargo, esta fase no se considerará a la hora de realizar este proyecto.

Como se puede comprobar a través de la figura anterior 3.15, las fases de Análisis, ETL y Evaluación conforman un ciclo iterativo intermedio dentro del ciclo de vida completo que permite refinar cada una de estas fases de forma iterativa hasta llegar a consolidar una versión final adecuada a las necesidades del proyecto y que cumpla con los objetivos perseguidos con su realización.

3.3. Estimación

La estimación dentro de un proyecto de software es una tarea de gran importancia que debe realizarse como parte de su planificación para ayudar a determinar la organización temporal del proyecto, el coste económico asociado, los recursos que será necesarios invertir para su desarrollo y, por supuesto, la viabilidad del proyecto en tiempo y forma.

En este apartado se hará una estimación a alto nivel del esfuerzo necesario para completar las historias de usuario que se plantean dentro del alcance del proyecto. En la planificación de cada sprint del proyecto, estas historias de usuario se desglosan en tareas más reducidas, propuestas para abordar de forma paulatina su contenido de cara a satisfacer los criterios de aceptación que tienen establecidos. La estimación del esfuerzo es efectuada sobre cada una de las tareas planificadas en los sprints mediante Planning Poker [48], una técnica de estimación en la que se asigna a cada tarea una dificultad en Puntos de Historia siguiendo una escala de valores determinada por la serie de Fibonacci. Esta serie numérica es aquella en la que cada número que la compone se obtiene a través de la suma de los 2 números inmediatamente anteriores. Esta escala de valores es incremental, de forma que la dificultad estimada para la tarea se incrementa según aumenta el valor en Puntos de Historia asignados a la misma. Aunque la estimación se llevará a cabo sobre cada una de las tareas planificadas en los sprints del proyecto, en esta memoria solo se indicarán los puntos de historia agrupados por historia de usuario, siendo su valor de estimación del esfuerzo igual a la suma de los Puntos de Historia asignados a cada una de las tareas en las que se dividen, con el objetivo de favorecer una visión más compacta y reducida de la tarea de estimación.

A continuación, se identifican y describen cada una de las historias de usuario comprendidas en el proyecto junto con su estimación del esfuerzo requerido en unidades de Puntos de Historia.

- **HU-01: ESTUDIO DEL ENTORNO DE NEGOCIO** (10 Puntos de Historia)

  Consiste en el estudio general del contexto del transporte urbano inteligente y, específicamente, del problema de construcción de matrices OD de tránsito. Además, incluye el estudio del conjunto de desafíos que surgen en torno a la construcción de matrices OD, tales como la estimación de la parada de bajada o la detección de transbordos. Igualmente, se trata la adaptación concreta del problema al caso de uso particular de la red de transporte público de Madrid.

- **HU-02: REVISIÓN DEL ESTADO DEL ARTE** (12 Puntos de Historia)

  Consiste en la revisión de las iniciativas de investigación anteriores que tratan el problema de construcción de matrices OD de tránsito para comprobar las soluciones propuestas en la literatura para abordar los desafíos involucrados y determinar posibles opciones que pueden adoptarse en el contexto de la red de transporte público de Madrid.

- **HU-03: BÚSQUEDA Y EXPLORACIÓN DE CONJUNTOS DE DATOS** (7 Puntos de Historia)

  Consiste en la búsqueda de potenciales conjuntos de datos que puedan considerarse para su uso en el proyecto. También, incluye la realización de una exploración inicial de sus contenidos, con el objetivo de determinar su adecuación a los propósitos del proyecto.

- **HU-04 ANÁLISIS Y TRANSFORMACIÓN DE DATOS** (18 Puntos de Historia)

  Consiste en analizar en profundidad los datos disponibles para conocer su estructura y contenido y así poder determinar las transformaciones sobre ellos y/o las integraciones con otros conjuntos de datos que pudieran ser necesarias para conformar unos datos acordes a los fines analíticos del proyecto. Esta historia de usuario incluye la realización del proceso ETL completo requerido para convertir el *raw data* disponible en el *smart data* refinado a evaluar.

- **HU-05: IMPLEMENTACIÓN DEL MÉTODO *TRIP CHAINING*** (17 Puntos de Historia)

  Consiste en realizar la codificación del método *trip chaining* designado para reconstruir los viajes realizados por los pasajeros de la red de transporte público de Madrid. Dentro de esta historia de usuario se incluye la propia detección de transbordos necesaria para determinar si un pasajero ha bajado del transporte público para llevar a cabo una actividad o para proseguir el viaje hasta su destino.

- **HU-06: VISUALIZACIÓN DE RESULTADOS** (12 Puntos de Historia)

  Consiste en el diseño e implementación de los gráficos que muestran los resultados obtenidos en el proyecto de cara a su inclusión en la memoria técnica. Dentro de esta historia de usuario se elaboran las matrices OD de tránsito y los mapas de soporte a la interpretación del funcionamiento de **trip chaining** y otros resultados significativos del proyecto. También, se disponen en tablas las estadísticas generales de viajes reconstruidos obtenidas tras la aplicación de **trip chaining**.

- **HU-07: DISCUSIÓN DE RESULTADOS** (5 Puntos de Historia)

  Consiste en la valoración crítica de los resultados finales obtenidos en el proyecto con el fin de determinar la viabilidad de la aplicación del método *trip chaining* para reconstruir los viajes realizados por los pasajeros de la red de transporte público de Madrid. Dentro de esta historia de usuario se incluyen las conclusiones y reflexiones finales sobre los aspectos positivos y negativos cuya corrección o mejora se podría realizar para optimizar la solución propuesta.

- **HU-08: ELABORACIÓN DE LA DOCUMENTACIÓN DEL PROYECTO** (13 Puntos de Historia)

Consiste en la redacción de la memoria técnica del proyecto, abarcando los diferentes capítulos y apéndices que la componen, junto con la elaboración del resto de entregables del proyecto.

- **HU-09: PREPARACIÓN DE LA PRESENTACIÓN** (9 Puntos de Historia)

  Consiste en la elaboración incremental de la presentación del proyecto que será utilizada posteriormente en el acto de defensa del mismo.

- **HU-10: IMPLEMENTACIÓN DE MEJORAS DEL INCREMENTO ANTERIOR** (3 Puntos de Historia)

  Consiste en la implementación, al inicio de cada sprint, de las mejoras derivadas de la retroalimentación aportada sobre el incremento entregado al finalizar el sprint anterior.

Finalmente, el alcance del proyecto se ha estimado en un total de 103 Puntos de Historia, cuya realización se llevará a cabo en forma de tareas a lo largo de los sprints que constituyen el proyecto.

3.4. Planificación Temporal

Tras haber especificado las historias de usuario que componen el alcance del proyecto, junto con su estimación del esfuerzo requerido, se procede a establecer la propuesta de planificación temporal a seguir durante la realización del proyecto para completar dichas historias.

Este proyecto se encuadra dentro de la asignatura Trabajo Fin de Máster (TFM), la cual lleva asociado un peso de 9 créditos ECTS, que equivalen a 225 horas de trabajo por parte del alumno. Por tanto, esta restricción temporal ha sido tenida en cuenta a la hora de planificar tanto el alcance del proyecto como la organización del trabajo del conjunto de sprints que lo componen.

Así, la planificación temporal del proyecto se ha organizado en 4 sprints de 3 semanas de duración cada uno, siguiendo la metolodogía de trabajo UVagile. El período empleado para su realización ha sido desde el inicio el día 14/06/2022 hasta su finalización el día 06/09/2022. De acuerdo con la dinámica de trabajo adoptado por UVagile, al finalizar cada sprint se conforma un incremento de producto, resultado de abordar tanto las historias de usuario específicas del sprint como las comunes a todos ellos, tales como la elaboración de la documentación (HU-08) y la preparación de la presentación (HU-09), además de las correcciones y mejoras surgidas a raíz de la retroalimentación recibida anteriormente (HU-10).

En la Tabla 3.12 se muestra la organización temporal de los sprints del proyecto junto con la asignación de historias de usuario en cada uno de ellos.

Sprint	Intervalo de Tiempo	Historias de Usuario comprendidas	Esfuerzo (Puntos de Historia)
#1	14/06/2022 - 05/07/2022	HU-01, HU-02, HU-08, HU-09	26
#2	05/07/2022 - 26/07/2022	HU-03, HU-04, HU-08, HU-09	29
#3	26/07/2022 - 16/08/2022	HU-05, HU-06, HU-08,	24
#4	16/08/2022 - 06/09/2022	HU-06, HU-07, HU-08, HU-09	24

Tabla 3.12: Planificación temporal de los sprints del proyecto

Igualmente, la planificación temporal del proyecto ha sido sintetizada a modo de diagrama de Gantt para facilitar su visualización (véase la Figura 3.16).

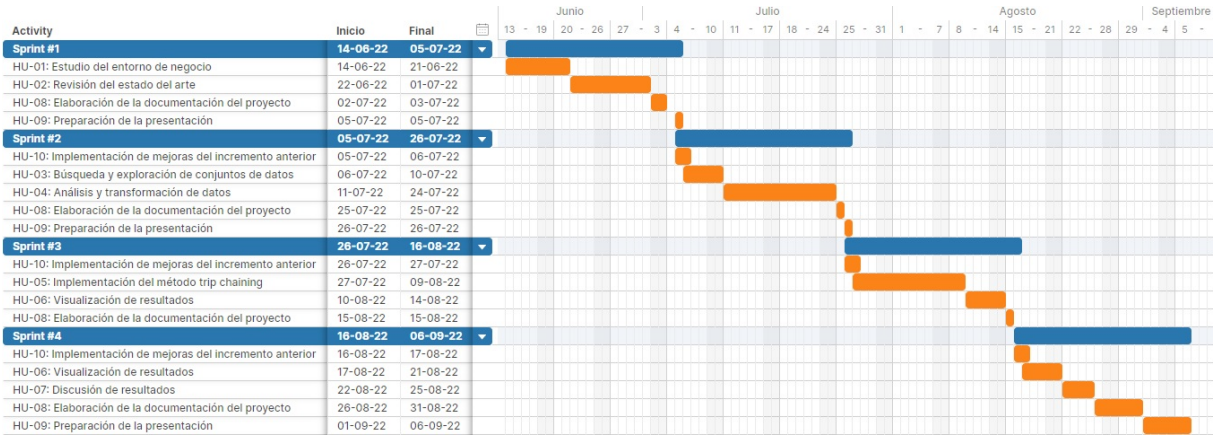

Figura 3.16: Diagrama de Gantt con la planificación temporal del proyecto.

3.5. Presupuesto

Una vez completada la planificación temporal del proyecto, la siguiente tarea será hacer una estimación del presupuesto económico necesario para poder llevarlo a cabo. Hay que tener en cuenta que el presupuesto es una estimación realizada a priori del coste que se prevé que conllevará la realización del proyecto y que, por tanto, el coste real al concluir el proyecto puede variar significativamente con respecto a lo previsto inicialmente.

Para la determinación del presupuesto asociado a un proyecto de software se deben considerar los costes derivados de los diversos medios necesarios para su ejecución, entre los

que destacan los costes de hardware, software y los honorarios de las personas encargadas de su elaboración.

A continuación, resumimos a través de tablas los costes estimados de los variados recursos involucrados en la ejecución del proyecto.

En primer lugar, desde el punto de vista del hardware, el coste asociado se deriva de los dispositivos a utilizar para el desarrollo del proyecto. En este caso, se considera el uso de un ordenador portátil de prestaciones medias, cuyas especificaciones técnicas más relevantes se indican en la Tabla 3.13.

Componente	**Prestación**
Procesador	Intel Core i7 - 7ª Generación
Tarjeta Gráfica	AMD Radeon R7 M440
Disco Duro	480 GB de SSD
Memoria RAM	8 GB

Tabla 3.13: Prestaciones del ordenador portátil para el desarrollo del proyecto

El coste de adquisición de dicho ordenador portátil que puede ser imputado al proyecto de acuerdo con el tiempo ue se ha dedicado al mismo se muestra en la Tabla 3.14. Este coste proporcional resultante se ha calculado por medio de la técnica del prorrateo, considerando un total de 85 días de dedicación al proyecto.

Ítem	**Precio**	**Tiempo de Vida Útil**	**Tiempo Proyecto**	**Uso en el proyecto**	**Prorrateo / Coste**
Ordenador Portátil	650 €	5 años (1826 días)	85 días	85*100/1826 = 4,65 %	650*85/1826 = **30,26 €**
TOTAL	**30,26 €**				

Tabla 3.14: Costes de Hardware

Por otro lado, dentro de los costes de software se incluyen principalmente los asociados a licencias de sistemas operativos, programas y herramientas utilizadas. Aunque el sistema operativo instalado en el equipo portátil es Windows 10, al llevar a cabo el desarrollo sobre una Máquina Virtual con distribución Ubuntu de Linux, el coste asociado a la licencia de Windows 10 no se considera dentro de los costes del presupuesto. Igualmente, los programas y herramientas empleadas en el proyecto son todas de código abierto o bien de versión gratuita, por lo que no repercuten en el presupuesto. Así, el desglose de los costes de software asociados al proyecto se indican en la Tabla 3.15.

A continuación, según la planificación temporal planteada, se calculan los costes relativos a los recursos humanos requeridos en el desarrollo del proyecto. Para ello, se

Ítem	Licencia	Coste
Microsoft Teams	Versión Gratuita	0 €
Trello	Libre	0 €
Anaconda	Libre	0 €
OpenRefine	Libre	0 €
TeXstudio	Libre	0 €
Draw.io	Libre	0 €
Creately	Versión Gratuita	0 €
Tom's Planner	Versión Gratuita	0 €
TOTAL	**0 €**	

Tabla 3.15: Coste de Software

deben tener en cuenta los diferentes roles de profesionales implicados. En lo que a este proyecto se refiere, se consideran necesarios los roles de Jefe de Proyecto, Analista de Datos y Desarrollador Big Data. Para la estimación del salario de cada uno de estos perfiles de profesionales, se ha consultado el sitio web de LinkedIn Salary[1], el cual proporciona datos acerca del salario medio según el puesto desempeñado. En la Tabla 3.16 se resumen los costes asociados a los recursos humanos, estimando las horas que cada uno de los roles debe invertir en la realización del proyecto.

Recurso	Salario	Horas	Coste
Jefe de Proyecto	4500 € / mes (26,80 € / hora)	45	26,80*45 = 1206 €
Analista de Datos	2750 € / mes (16,40 € / hora)	90	16,40*90 = 1476 €
Desarrollador Big Data	2750 € / mes (16,40 € / hora)	90	16,40*90 = 1476 €
TOTAL	**4158 €**		

Tabla 3.16: Coste de Recursos Humanos

Nota: para la obtención del salario diario se ha considerado una jornada laboral estándar de 21 días laborables y 8 horas diarias.

Por último, se considera un conjunto de Otros Costes entre los que se incluyen costes varios: la Conexión a Internet, la electricidad, el material de oficina, etc. En este caso, se limita el cálculo de los Otros Costes a la Conexión a Internet y la electricidad, tal y como se muestra en la Tabla 3.17.

[1] https://www.linkedin.com/salary/

Ítem	**Coste Mensual**	**Tiempo Proyecto**	**Prorrateo / Coste**
Conexión a Internet	50 €	85 días (2,75 meses)	2,75*50 = 137,50 €
Electricidad	15 €	85 días (2,75 meses)	2,75*15 = 41,25 €
TOTAL	**178,75 €**		

Tabla 3.17: Otros Costes

Una vez estimados los diferentes tipos de costes asociados al proyecto, se procede a calcular el coste total para obtener el presupuesto económico final (véase la Tabla 3.18).

Recurso	**Coste**
Hardware	30,26 €
Software	0 €
Humanos	4158 €
Otros	178,75 €
TOTAL	**4367,01 €**

Tabla 3.18: Presupuesto del Proyecto

Análisis

Teniendo en cuenta que la etapa de Inicio del proyecto ya ha sido completada en capítulos anteriores, la siguiente a realizar será la de Análisis, la cual constituye la primera etapa dentro del ciclo iterativo establecido en la metodología de desarrollo adoptada para refinar incrementalmente los datos y resultados del proyecto.

La etapa de Análisis de un proyecto *Big Data* comprende una serie de actividades preliminares de exploración, estudio y modelado de los datos que conforman una descripción completa de los mismos con el propósito de poder consolidar una primera abstracción del proyecto, la cual podrá ser actualizada convenientemente a medida que vayamos ampliando nuestro conocimiento sobre los datos durante la realización del resto de las etapas. Gracias a este análisis se dispondrá de una planificación inicial útil de cara a comenzar con la siguiente fase del ciclo de vida.

4.1. Modelado de Dominio

El modelado de dominio es la primera actividad dentro del análisis en la que se lleva a cabo una descripción, a distintos niveles de detalle, de los datos finales que compondrán la solución de nuestro proyecto, es decir, se encarga de planear y caracterizar el *smart data*. Este *smart data* se obtendrá a través de la realización de extensas modificaciones sobre los datos de los que se dispone inicialmente (*raw data*) y constituirá uno de los resultados fundamentales del proyecto.

4.1.1. Modelo conceptual

El modelo conceptual de datos plantea una propuesta de modelado del *smart data* a través de la creación de un conjunto de entidades de información y una serie de relaciones

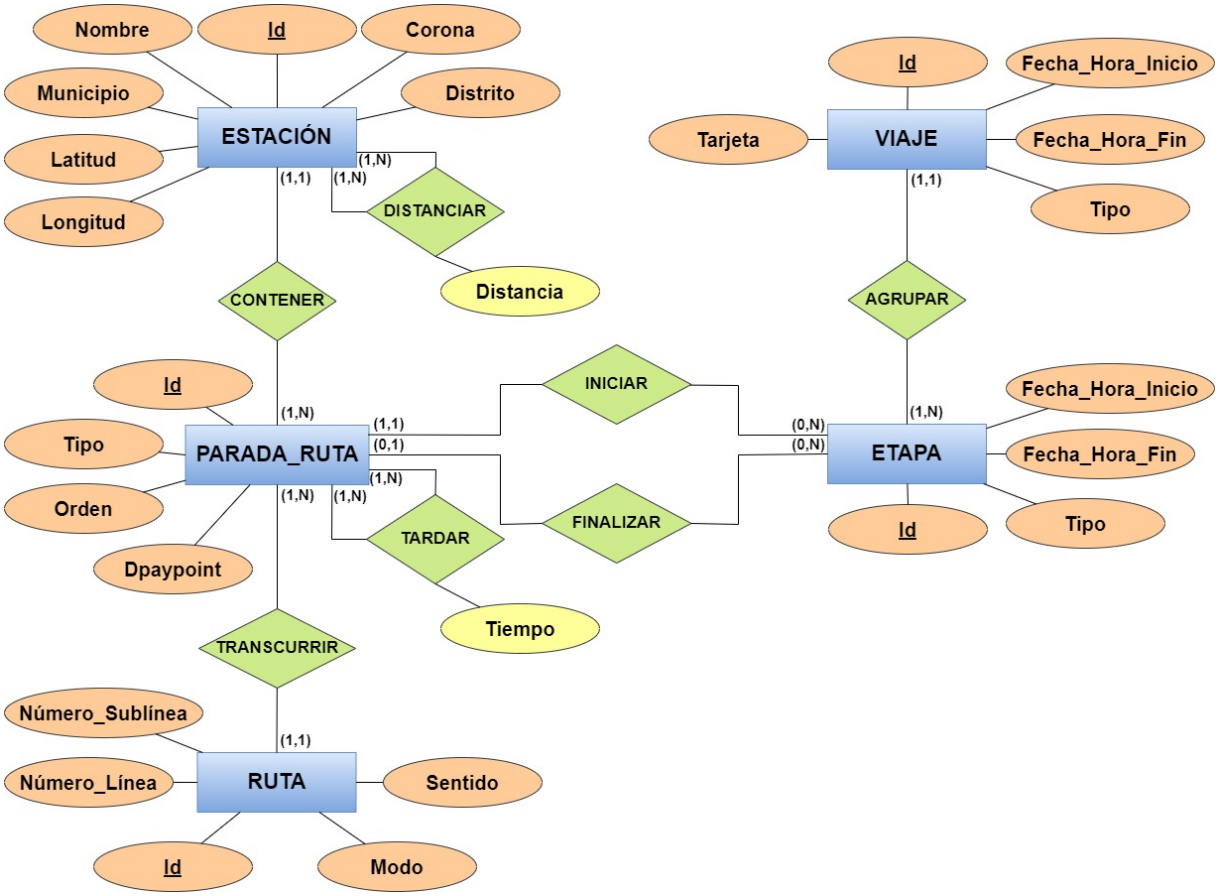

Figura 4.17: Diagrama Entidad-Relación.

que interconectan estas entidades entre sí. El propósito de esta actividad es generar una descripción de los datos acorde con la realidad que se desea modelar y que facilite una posterior Evaluación de acuerdo con los objetivos del proyecto.

En este caso, se elabora un diagrama entidad-relación (ER) para describir nuestro modelo conceptual de datos (véase la Figura 4.17). A continuación, se procede a explicar cada uno de los elementos descritos en el marco de este diagrama conceptual.

Entidades Las entidades de datos se corresponden con los principales tipos de información que intervienen en nuestro proyecto y están caracterizados por un conjunto de atributos que los describen.

- **ESTACIÓN**: representa las ubicaciones dentro de la red de transporte público donde los pasajeros se pueden subir y bajar a los vehículos de transporte en el transcurso de sus viajes.

- **RUTA**: representa los trayectos definidos en la red de transporte público que se ofrecen a los pasajeros para realizar sus viajes.
- **PARADA_RUTA**: representa las paradas por las que transcurren las rutas de transporte. Dentro de una misma estación puede haber paradas de varias rutas o de la misma ruta, pero en diferentes sentidos.
- **ETAPA**: representa los segmentos de viaje de los pasajeros, descritos por una parada de subida y otra de bajada.
- **VIAJE**: representa las agrupaciones de etapas con un origen y un destino de viaje definidos según el propósito de viaje de los pasajeros.

Relaciones Las relaciones asocian las distintas entidades de datos con el objetivo de modelar la existencia de una cierta conexión entre ellas.

- **CONTENER**: relaciona las paradas de ruta con la estación en la que se encuentran.
- **TRANSCURRIR**: relaciona las paradas de ruta con la ruta de transporte de la que forman parte.
- **INICIAR**: relaciona las etapas de viaje con su parada de subida correspondiente.
- **FINALIZAR**: relaciona las etapas de viaje con su parada de bajada correspondiente.
- **AGRUPAR**: relaciona las etapas de viaje con el viaje en el que se incluyen.
- **DISTANCIAR**: relaciona, en términos de distancia, pares de estaciones dentro de la red de transporte.
- **TARDAR**: relaciona, en términos de tiempo, pares de paradas de ruta dentro de sus rutas de transporte.

4.1.2. Diccionario de Datos

El diccionario de datos aporta una descripción de los datos más detallada que el modelo ER, especificando de forma pormenorizada las características de cada uno de los elementos de datos modelados. Éste constituye una fuente de metadatos que serán de utilidad para la realización de las siguientes fases del proyecto.

En este caso, se confecciona el diccionario de datos a modo de tablas con una estructura definida en las que se definen cada una de las entidades y relaciones descritas en el modelo conceptual junto con las propiedades de sus atributos descriptivos. Una tabla de ejemplo

ESTACIÓN						
Definición	Representa cada una de las ubicaciones dentro de la red de transporte público donde los pasajeros pueden subir y bajar de los vehículos de transporte.					
Notas	Desde una misma estación se pueden seguir distintas rutas de transporte.					
Atributos	**Nombre**	**Definición**	**Tipo**	**UNIQUE**	**NULL**	**DEFAULT**
	Id	Identificador de la estación en la red de transporte.	STRING	SÍ	NO	-
	Nombre	Denominación de la estación.	STRING	NO	NO	-
	Municipio	Municipio al que pertenece la estación.	STRING	NO	NO	-
	Distrito	Distrito en el que se encuentra la estación.	STRING	NO	SÍ	-
	Corona	Zona tarifaria en la que se ubica la estación.	ENUM	NO	SÍ	-
	Latitud	Coordenada de la latitud donde se ubica la estación expresada en grados.	FLOAT	NO	NO	-
	Longitud	Coordenada de la longitud donde se ubica la estación expresada en grados.	FLOAT	NO	NO	-

Figura 4.18: Tabla de ejemplo de entidad del Diccionario de Datos.

de entidad del diccionario de datos correspondiente a la entidad ESTACIÓN se muestra en la Figura 4.18. En ella, aparece una definición formal del término al que hace referencia la entidad, unas notas que proporcionan información adicional de interés acerca de la misma y una descripción detallada de cada uno de los atributos que forman parte de ella.

Por otro lado, en la Figura 4.19 se muestra un ejemplo de tabla de relación del diccionario de datos correspondiente a la relación TRANSCURRIR. La información presentada para este tipo de información es semejante al caso anterior, incluyendo además detalles acerca de la participación y cardinalidad propias de la relación establecida entre las entidades.

TRANSCURRIR			
Definición	Relaciona cada ruta de transporte con las paradas por las que transcurre.		
Notas			
Entidades	**Nombre**	**Participación**	**Cardinalidad**
	RUTA	1	N
	PARADA_RUTA	1	1

Figura 4.19: Tabla de ejemplo de relación del Diccionario de Datos.

El resto de tablas del diccionario de datos pueden ser consultadas en el Apéndice A.

4.2. Exploración de Datos

La exploración de datos se encarga de la identificación y estudio individualizado de potenciales conjuntos de datos para utilizar en el proyecto. El principal objetivo de esta actividad es ampliar el conocimiento sobre los datos disponibles, el denominado *raw data*, de cara a ir perfilando los tipos de datos que tendremos realmente a nuestra disposición para componer nuestro *smart data*.

4.2.1. Perfilado de Datos

El perfilado de datos, también denominado *data profiling*, es una actividad que consiste en la revisión y entendimiento del contenido y estructura de conjuntos de datos que han sido identificados previamente a la realización de alguna acción que los modifique, es decir, conjuntos de datos tal como han sido recopilados.

Entre la información más importante a analizar destacan características generales y de definición de los propios conjuntos de datos, así como la descripción de propiedades relevantes acerca de los atributos que los componen. De esta manera, se puede comprobar la calidad y corrección de los datos y proponer posibles medidas de actuación que puedan ser necesarias para adaptar los datos a su uso en el proyecto. Finalmente, la motivación de realizar este proceso es determinar la viabilidad en el uso de los conjuntos de datos para la realización del proyecto por medio del estudio de la validez y adecuación de los datos a los objetivos que se necesitan abordar en el proyecto.

En este caso, se han confeccionado diferentes tablas de perfiles de datos que resumen el estudio de exploración realizado sobre los conjuntos de datos. En la Tabla 4.19 se especifican los conjuntos de datos analizados junto con su información de procedencia. Hay que tener en cuenta que para la realización de este proyecto se priorizan las fuentes de datos abiertos con el objetivo de completar el proceso con la mayor cantidad de datos que puedan estar a disposición de cualquier persona interesada.

Tabla 4.19: Conjuntos de Datos recopilados

Conjunto de Datos	**Contenido**	**Procedencia**
M4 ESTACIONES	Información descriptiva de las estaciones de Metro de la red de transporte público de Madrid.	Portal de Datos Abiertos del CRTM (https://data-crtm.opendata.arcgis.com/maps/m4-estaciones). Actualizado a 30 de noviembre de 2018.
M4 TRAMOS	Información descriptiva de los diferentes tramos que componen cada una de las rutas de transporte ofrecidas en el Metro.	Portal de Datos Abiertos del CRTM (https://data-crtm.opendata.arcgis.com/maps/m4-tramos). Actualizado a 30 de noviembre de 2018.
M10 ESTACIONES	Información descriptiva de las estaciones de Metro Ligero de la red de transporte público de Madrid.	Portal de Datos Abiertos del CRTM (https://data-crtm.opendata.arcgis.com/maps/m10-estaciones). Actualizado a 22 de junio de 2016.
M10 TRAMOS	Información descriptiva de los diferentes tramos que componen cada una de las rutas de transporte ofrecidas en el Metro Ligero.	Portal de Datos Abiertos del CRTM (https://data-crtm.opendata.arcgis.com/maps/m10-tramos). Actualizado a 22 de junio de 2016.
M5 ESTACIONES	Información descriptiva de las estaciones de Cercanías de la red de transporte público de Madrid.	Portal de Datos Abiertos del CRTM (https://data-crtm.opendata.arcgis.com/maps/m5-estaciones). Actualizado a 25 de septiembre de 2019.

Continúa en la página siguiente

Tabla 4.19 – continuación de la página anterior

Conjunto de Datos	Contenido	Procedencia
M5 TRAMOS	Información descriptiva de los diferentes tramos que componen cada una de las rutas de transporte ofrecidas en Cercanías.	Portal de Datos Abiertos del CRTM (https://data-crtm.opendata.arcgis.com/maps/m5-tramos). Actualizado a 25 de septiembre de 2019.
M6 ESTACIONES	Información descriptiva de las paradas de Autobuses EMT de la red de transporte público de Madrid.	Portal de Datos Abiertos del CRTM (https://data-crtm.opendata.arcgis.com/maps/m6-estaciones). Actualizado a 12 de septiembre de 2019.
M6 TRAMOS	Información descriptiva de los diferentes tramos que componen cada una de las rutas de transporte ofrecidas en Autobuses EMT.	Portal de Datos Abiertos del CRTM (https://data-crtm.opendata.arcgis.com/maps/m6-tramos). Actualizado a 12 de marzo de 2021.
M8 ESTACIONES	Información descriptiva de las paradas de Autobuses Interurbanos de la red de transporte público de Madrid.	Portal de Datos Abiertos del CRTM (https://data-crtm.opendata.arcgis.com/maps/m8-estaciones). Actualizado a 12 de septiembre de 2019.
M8 TRAMOS	Información descriptiva de los diferentes tramos que componen cada una de las rutas de transporte ofrecidas en Autobuses Interurbanos.	Portal de Datos Abiertos del CRTM (https://data-crtm.opendata.arcgis.com/maps/m8-tramos). Actualizado a 12 de septiembre de 2019.

Continúa en la página siguiente

Tabla 4.19 – continuación de la página anterior

Conjunto de Datos	Contenido	Procedencia
TOPOLOGÍA TRENES	Información topológica e identificativa de las estaciones de Metro, Metro Ligero y Cercanías de la red de transporte público de Madrid.	Petición al Consorcio Regional de Transportes de Madrid. Actualizado a 29 de febrero de 2020.
TOPOLOGÍA EMT	Información topológica e identificativa de las paradas de Autobuses EMT de la red de transporte público de Madrid.	Petición al Consorcio Regional de Transportes de Madrid. Actualizado a 29 de febrero de 2020.
TOPOLOGÍA INTERURBANOS	Información topológica e identificativa de las paradas de Autobuses Interurbanos de la red de transporte público de Madrid.	Petición al Consorcio Regional de Transportes de Madrid. Actualizado a 29 de febrero de 2020.
ENERO VIAJES 2019	Transacciones de viaje realizadas en el mes de enero de 2019 en el transporte público de Madrid.	Petición al Consorcio Regional de Transportes de Madrid. Actualizado a 31 de enero de 2019.

A continuación, se presentan a modo de ejemplo las tablas de perfil de datos asociadas al conjunto Enero Viajes 2019 proporcionado por el CRTM.

Primero, la Figura 4.20 establece las consideraciones generales sobre el conjunto de datos, incluyendo información adicional, como por ejemplo las anomalías detectadas y sus posibles vías de solución. Por su parte, la Figura 4.21 describe cada uno de los atributos de datos de cara a su caracterización preliminar. Por último, la Figura 4.22 muestra un análisis más detallado acerca del contenido de los datos, estudiando características como la cardinalidad, la completitud, el formato de valores o diversas propiedades numéricas, entre las que se encuentran los valores máximo y mínimo, el valor media o la desviación estándar.

El resto de tablas de perfiles de datos pueden ser consultadas en el Apéndice B.

ENERO VIAJES 2019	
Definición	Contiene las transacciones de viaje realizadas en el mes de enero de 2019 en el transporte público de Madrid.
Fuente	Petición al Consorcio Regional de Transportes de Madrid.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero TXT (campos separados por "#").
Crecimiento	Conjunto de datos dinámico.
Fecha de Actualización	31 de enero de 2019.
Anomalías	- Presencia de valores en el campo DPAYPOINT que no aparecen en los conjuntos de datos estáticos de la red de transporte público de Madrid.
Decisiones	- Descartar las transacciones con valores de DPAYPOINT no reconocibles, así como el resto de las transacciones realizadas con la misma tarjeta de pasajero en el día en cuestión.

Figura 4.20: Tabla Primera del perfil de datos de Enero Viajes 2019.

ENERO VIAJES 2019		
Campo	**Descripción**	**Tipo de Datos**
TARJETA	Hash identificador de la tarjeta de transporte de un pasajero de la red de transporte público de Madrid.	STRING
FECHA	Fecha y hora correspondientes a la transacción de viaje.	DATETIME
TUSUARIO	Código asociado al perfil de usuario de la tarjeta de transporte.	STRING
TITULO	Código asociado al título de transporte del viaje.	STRING
DESCUENTO	Código de descuento aplicado al viaje.	STRING
DPAYPOINT	Código identificativo de los terminales de registro de transacciones asociados a la parada donde se ha realizado la transacción. Está formado por el código del operador, el número de la línea y el número de la parada.	STRING
IDTLV	Código indicador del modo de registro y la procedencia de la transacción.	STRING
CODVAL	Código de validación relativo al resultado y tipología de la transacción.	STRING
Otros atributos: -		

Figura 4.21: Tabla Segunda del perfil de datos de Enero Viajes 2019.

ENERO VIAJES 2019							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
TARJETA	4588825	100%	-	-	-	-	[A-Z0-9]{32}
FECHA	1212556	100%	01/01/2019 00:00:04	31/01/2019 23:59:58	-	-	2019-01-(0[1-9]|1[0-9]|2[0-9]|3[01]) (0[1-9]|1[0-9]|2[0-4]):[0-5][0-9]:[0-5][0-9]
TUSUARIO	10	100%	-	-	-	-	01|02|03|04|06|07|08|0B|0D|0F
TITULO	147	100%	-	-	-	-	[A-F0-9]{4}
DESCUENTO	6	99,85%	00	05	-	-	00|01|02|03|04|05
DPAYPOINT	31399	100%	-	-	-	-	[A-F0-9]{2}_L[0-9]{1,4}_P[0-9]{1,5}
IDTLV	4	100%	-	-	-	-	C0|C6|C1|G2
CODVAL	117	100%	0	177	-	-	[0-9]{1,3}
Número Total de Registros							151.048.520

Figura 4.22: Tabla Tercera del perfil de datos de Enero Viajes 2019.

4.2.2. Análisis de los datos de transacciones de viaje

A continuación, se repasan de forma más pormenorizada características cuantitativas y estructurales acerca de los datos de transacciones de viaje que han sido puestos a nuestros disposición por el CRTM. Para ello, en las siguientes tablas se revisan diferentes agregaciones de los datos de transacciones que proporcionan un mayor entendimiento a la hora de conocer en profundidad estos datos e inspeccionar sus posibilidades de uso.

En primer lugar, en la Tabla 4.20 se presenta un análisis cuantitativo del número de transacciones que pueden ser agrupadas en torno a los distintos tipos de perfiles de usuario que son gestionados en la red de transporte público de Madrid. De manera similar, las Tablas 4.21 y 4.22 muestran una agrupación de las transacciones por tipo de título tarifario involucrado y descuento aplicado, respectivamente. Por otra parte, la Tabla 4.23 especifica estadísticas de resumen del conjunto de transacciones junto con las causas principales de descarte de algunas de ellas. La explicación a dichas causas de descarte de las transacciones de viaje puede consultarse en la Tabla 4.24. Como estrategia de descarte de transacciones, se ha optado por eliminar todas las transacciones asociadas a una tarjeta de transporte en el día considerado si alguna de ellas presenta alguna anomalía, al considerar que dejando las transacciones restantes se produciría una interrupción de la cadena de viajes asociadas al pasajero, que infringiría los principios del método *trip chaining* a aplicar posteriormente.

Perfil de Usuario	**Número de Transacciones**
Normal	88.515.660
Joven	40.908.676
3ª Edad	17.530.422
Tarjeta Azul	2.528.149
Infantil	1.189.378
Turístico Normal	363.253
Turístico Infantil	11.649
Programa Activación y Empleo	1.130
Pensionista Valdemoro	136
Discapacitado 33 % Valdemoro	48

Tabla 4.20: Transacciones de viaje por perfil de usuario

Tipo de Título Tarifario	**Número de Transacciones**
Abono 30 Días	119.661.653
Billete 10 Viajes	19.427.481
Billete Sencillo	5.524.463
Abono Anual	5.014.796
Abono Infantil	1.189.360
Abono Autobuses Interurbanos	852.832
Abono Autobuses Urbanos	792.239
Abono Turístico	382.959
Abono Empleados (No Pasajeros)	137.329
Abono Municipal Valdemoro	185

Tabla 4.21: Transacciones de viaje por tipo de título tarifario.

4.3. Diseño del *Pipeline* de Transformación

El diseño del *pipeline* de transformación de datos es una actividad llevada a cabo para planificar la transformación de los datos disponibles inicialmente (*raw data*) en los datos finales (*smart data*) requeridos en la etapa de Evaluación del proyecto.

Antes de proceder con la definición de este *pipeline* debemos saber cómo estará organizado el *smart data* a obtener en el proyecto.

Descuento Aplicado	**Número de Transacciones**
Sin Descuento	141.959.865
Familia Numerosa General	6.714.131
Familia Numerosa Especial	1.250.246
Discapacidad >= 65 %	773.989
Discapacidad >= 65 + Familia Numerosa General	96.692
Discapacidad >= 65 + Familia Numerosa Especial	21.389

Tabla 4.22: Transacciones de viaje por descuento aplicado.

Nº Transacciones Inicial	**151.048.520**
Nº Total de Transacciones a descartar	**18.812.844**
- Transacciones únicas en un día	150.302
- Dpaypoint no reconocido	4.576.322
- Código de validación erróneo	225.601
- Código de validación no confirmado	201.191
- Relacionadas con las anteriores (tarjeta y día coincidentes)	13.421.790
- Tarjetas de empleados de la red de transporte	137.329
- Tarjetas de excepciones	100.309
Nº Transacciones Final	**132.235.676** (72 % del total)

Tabla 4.23: Estadísticas Resumen con las causas de descarte de transacciones de viaje.

4.3.1. Modelo Lógico de Datos

El modelo lógico de datos permite representar la organización de los datos de forma independiente al sistema de almacenamiento físico a utilizar, sin considerar suposiciones a bajo nivel de implementación. De esta manera, constituye un mecanismo efectivo para modelar la estructura de los elementos de datos y las relaciones que existen entre ellos.

A continuación, en la Figura 4.23 se muestra el modelo relacional diseñado como modelo lógico de datos para describir la organización del *smart data*. En ella, se puede observar la definición en forma tabular de las diferentes entidades de datos y relaciones que aparecen descritas en el modelo conceptual de datos de la Figura 4.17.

Causa Descarte de Transacciones	Descripción
Transacciones únicas en un día	Correspondiente a tarjetas que solo tienen una transacción en el día considerado.
Dpaypoint no reconocido	Correspondiente a transacciones con un dpaypoint no localizado en los datos disponibles.
Código de validación erróneo	Correspondiente a transacciones con un código de validación de error (CODVAL=0).
Código de validación no reconocido	Correspondiente a transacciones con un código de validación no localizado en los datos disponibles.
Relacionadas con las anteriores	Correspondiente a transacciones cuya tarjeta y día asociados son iguales que los de alguna de las transacciones descartadas por una de las causas anteriores.
Tarjetas de empleados	Correspondiente a tarjetas de transporte de empleados de la red de transporte público de Madrid.
Tarjetas de excepciones	Correspondiente a tarjetas de transporte cuya frecuencia de transacciones supera el umbral de 20 transacciones de viaje en un mismo día virtual y, por tanto, no representan los patrones de viaje de los pasajeros comunes.

Tabla 4.24: Explicación de las causas de descarte de transacciones de viaje.

4.3.2. *Pipeline* de transformación de datos

El *pipeline* de transformación de datos (*dataflow*) es un artefacto diseñado para especificar de forma general la descomposición de tareas de transformación de datos que se plantean para acometer la conversión del *raw data* en *smart data*.

A continuación, en la Figura 4.24 se presenta el *pipeline* de transformación de datos diseñado. En él se muestra la descomposición en una serie de tareas del proceso de transformación de datos que se deberá implementar durante la realización del proceso ETL. Por un lado, en la parte izquierda del diagrama, se representan los conjuntos de datos iniciales (*raw data*), los cuales, tras aplicar sobre ellos las tareas de transformación diseñadas, darán lugar a los conjuntos de datos finales (*smart data*), que pueden verse en la parte derecha del diagrama. El proceso de transformación a realizar es sintetizado a través de las diferentes tareas situadas entre ambos tipos de conjuntos de datos, las cuales son planificadas para su implementación secuencial en distintos momentos del proceso general. Estas tareas dan como resultado nuevos conjuntos de datos modificados de acuerdo con la lógica implementada en cada una.

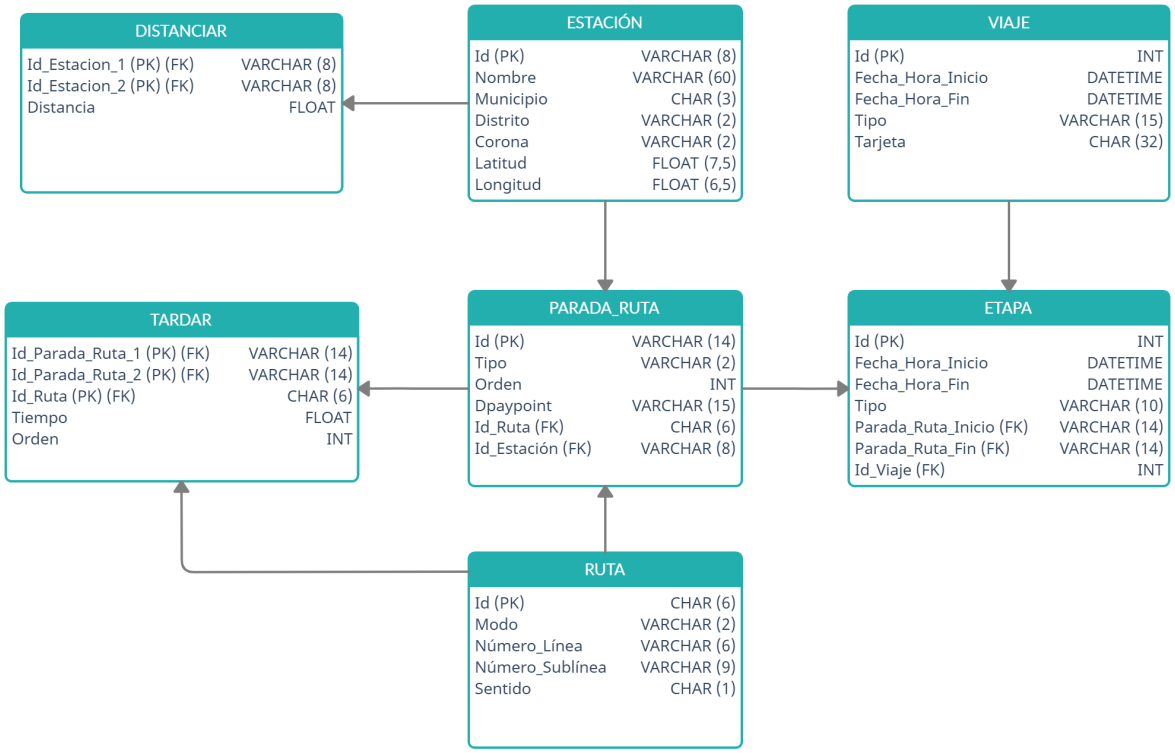

Figura 4.23: Modelo Lógico de Datos.

4.3.3. *Logical Datamap*

El *logical datamap* es una herramienta empleada para desglosar el proceso de transformación de datos en torno a cada uno de los conjuntos de datos. De esta forma, es posible dar una visión de la contribución a la realización de la tarea de cada uno de los atributos que componen los diferentes conjuntos de datos.

En este caso, el *logical datamap* es especificado a través de tablas correspondientes a cada uno de los conjunto de datos que se modifican durante el proceso de transformación. A modo de ejemplo, en la Tabla 4.25 se muestra la tabla del *logical datamap* correspondiente al conjunto de datos Enero Viajes 2019. En ella, aparece la contribución de los distintos atributos de este conjunto de datos a la realización de cada una de las tareas de transformación. Nótese que, para facilitar su visualización, las primeras tareas del *pipeline* de transformación de datos han sido omitidas al no utilizarse este conjunto de datos en ninguna de ellas. En este caso, tan sólo las tareas tripChaining y agruparEtapas requieren atributos de este conjunto de datos para su procesamiento.

El resto de tablas que constituyen el *logical datamap* completo pueden ser consultadas en el Apéndice C.

(14) ENERO VIAJES 2019	2.1. preprocesar-Transacciones	4.1. tripChaining	5.1. agruparEtapas	**CONJUNTO RESULTADO**
IdEtapa	PROCESO	-	-	**ETAPAS**
Tarjeta	PROCESO	PROCESO	PROCESO	**VIAJES**
Fecha	PROCESO	PROCESO	PROCESO	**ETAPAS y VIAJES**
TUsuario	PROCESO	✗		
Titulo	PROCESO	✗		
Descuento	PROCESO	✗		
Dpaypoint	PROCESO	JOIN con PARADAS RUTAS	-	-
IdTlv	PROCESO	✗		
CodVal	PROCESO	PROCESO	-	-
		Creación de ParadaSubida	-	**ETAPAS**
		Creación de ParadaBajada	-	**ETAPAS**
		Creación de FechaHoraFin	-	**ETAPAS y VIAJES**
			Creación de IdViaje	**ETAPAS y VIAJES**

Tabla 4.25: Tabla del *logical datamap* correspondiente al conjunto Enero Viajes 2019.

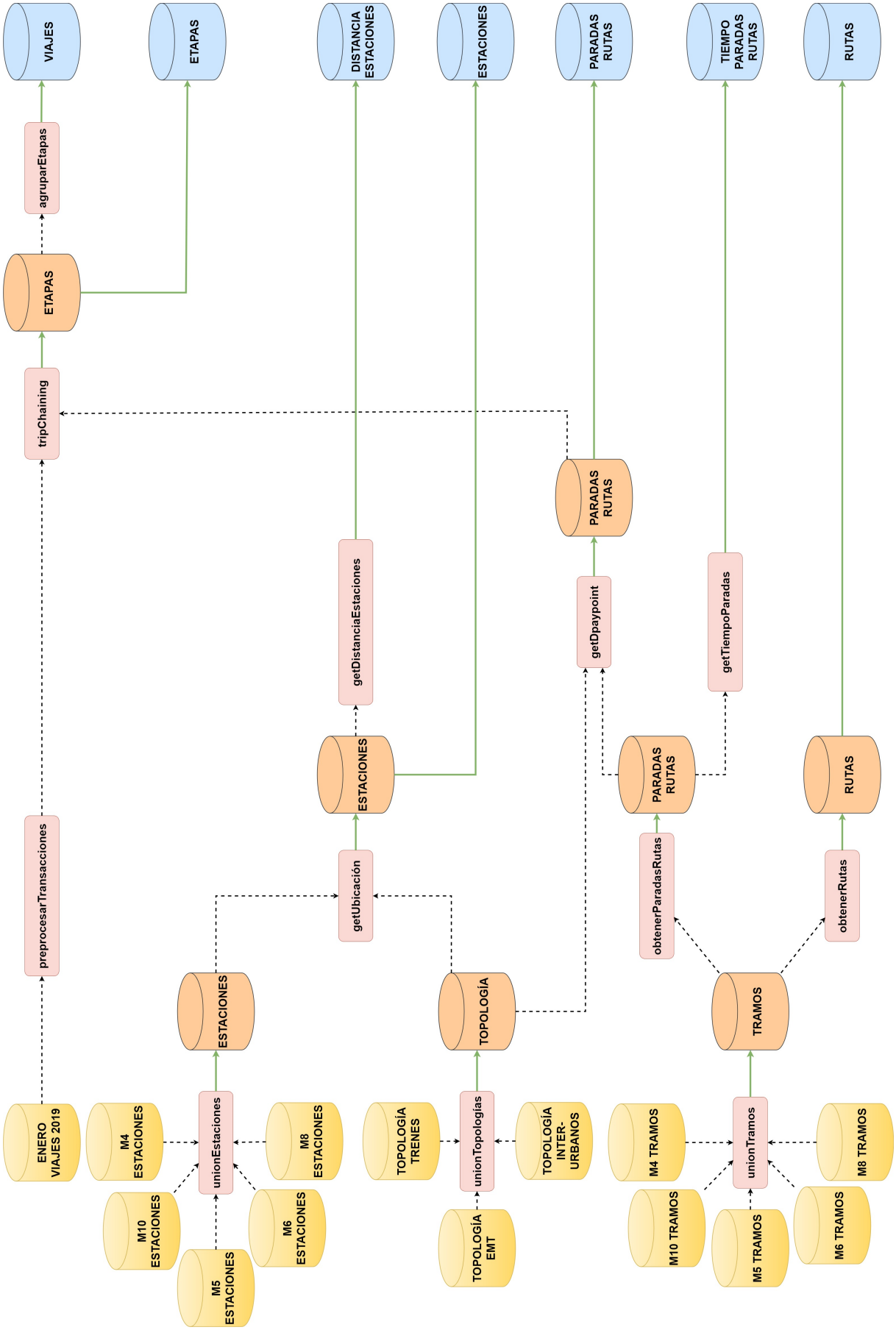

Figura 4.24: Pipeline de transformación de datos (*Dataflow*).

Proceso ETL

La siguiente etapa a realizar en el proyecto, de acuerdo con la metodología de desarrollo adoptada, se corresponde con el proceso ETL (Extracción, Transformación y Carga), a través del del cual se convierten los datos iniciales disponibles (*raw data*) en datos refinados (*smart data*) para su uso efectivo en el proyecto. En este caso, las operaciones de transformación acometidas sobre los datos estarán dirigidas a adecuar los datos para la aplicación del método *trip chaining*.

5.1. Extracción

En todo proceso ETL el primer paso consiste en realizar la extracción de los conjuntos datos desde sus fuentes de origen. En este caso, una parte de los conjuntos de datos son descargados directamente del Portal Web de Datos Abiertos del CRTM sin necesidad de especificar ningún tipo de clave o *token*, al ser datos abiertos de libre acceso para cualquier persona. Mientras, el resto de conjuntos de datos que han sido proporcionados por el CRTM son accedidos a través de un volumen de disco duro virtual.

Una vez descargados los datos, éstos serán cargados para su lectura y manejo desde el programa de manipulación de datos utilizado. En este caso, utilizamos Python, junto con la librería Pandas, para hacer la carga de los conjuntos de datos sobre *Dataframes*, una estructura de datos de 2 dimensiones en la que los registros se disponen en filas y sus atributos correspondientes en columnas.

A continuación, se presenta un registro de muestra por conjunto de datos con el objetivo de ilustrar la estructura que tienen los conjuntos de datos iniciales a transformar durante este proceso. Se debe tener en cuenta que las muestras de datos que se exponen representan solo el conjunto parcial de las columnas que han sido leídas para su posterior procesamiento.

Conjuntos de Datos de Estaciones. Agrupa los conjuntos de datos con la información acerca de las estaciones donde pueden subir y bajar los pasajeros de la red de transporte público de Madrid.

- **M4 ESTACIONES**: representa las estaciones de Metro de la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.26.

IDESTACION	DENOMINACION	CODIGOMUNICIPIO	DISTRITO	CORONATARIFARIA	X	Y
4_38	NOVICIADO	079	01	A	440100	4475360

Tabla 5.26: Registro de muestra del conjunto de datos M4 ESTACIONES.

- **M10 ESTACIONES**: representa las estaciones de Metro Ligero de la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.27.

IDESTACION	DENOMINACION	CODIGOMUNICIPIO	DISTRITO	CORONATARIFARIA	X	Y
10_50	PLAZA DE TOROS	106	01	B2	435675	4455333

Tabla 5.27: Registro de muestra del conjunto de datos M10 ESTACIONES.

- **M5 ESTACIONES**: representa las estaciones de Cercanías de la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.28.

IDESTACION	DENOMINACION	CODIGOMUNICIPIO	DISTRITO	CORONATARIFARIA	X	Y
5_70	SAN CRISTÓBAL INDUSTRIAL	079	17	A	440757	4465039

Tabla 5.28: Registro de muestra del conjunto de datos M5 ESTACIONES.

- **M6 ESTACIONES**: representa las paradas de Autobuses EMT de la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.29.

IDESTACION	DENOMINACION	CODIGOMUNICIPIO	DISTRITO	CORONATARIFARIA	X	Y
6_1514	OFELIA NIETO-FRANCOS RODRIGUEZ	079	09	A	439802	4479051

Tabla 5.29: Registro de muestra del conjunto de datos M6 ESTACIONES.

- **M8 ESTACIONES**: representa las paradas de Autobuses Interurbanos de la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.30.

IDESTACION	DENOMINACION	CODIGOMUNICIPIO	DISTRITO	CORONATARIFARIA	X	Y
8_12982	CTRA.M533-URB.EL ALCOR	160	01	C1	405845	4488920

Tabla 5.30: Registro de muestra del conjunto de datos M8 ESTACIONES.

Conjuntos de Datos de Topologías. Agrupa los conjuntos de datos con la información topológica acerca de las estaciones de la red de transporte público de Madrid.

- **TOPOLOGÍA TRENES**: representa la información topológica de las estaciones de Metro, Metro Ligero y Cercanías de la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.31.

IDFESTACION	LATITUD	LONGITUD	DPAYPOINT
4_24	40.38732	-3.63951	02_L1_P24

Tabla 5.31: Registro de muestra del conjunto de datos TOPOLOGÍA TRENES.

- **TOPOLOGÍA EMT**: representa la información topológica de las paradas de Autobuses EMT de la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.32.

IDFPARADAGESTRA	LATITUD	LONGITUD	NULINGES	NULINUSER	IDPARADAE	DPAYPOINT
6_31	40.42801	-3.71454	001	1	193	03_L1_P193

Tabla 5.32: Registro de muestra del conjunto de datos TOPOLOGÍA EMT.

- **TOPOLOGÍA INTERURBANOS**: representa la información topológica de las paradas de Autobuses Interurbanos de la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.33.

IDFPARADAGESTRA	LATITUD	LONGITUD	IDLINEA	NULINUSER	DPAYPOINT
8_16932	40.51606	-3.64211	2	2	30_L2_P254

Tabla 5.33: Registro de muestra del conjunto de datos TOPOLOGÍA INTERURBANOS.

Conjuntos de Datos de Tramos de Rutas. Agrupa los conjuntos de datos con la información acerca de los tramos de cada una de las rutas de transporte ofrecidas a los pasajeros de la red de transporte público de Madrid.

- **M4 TRAMOS**: representa los tramos de rutas de Metro ofrecidas en la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.34.

MODO	CODITINERARIO	NUMLINEAUSUARIO	SENTIDO	NUMORDEN	TIPOPARADA	LONGITUDTRAMOANT	VELOCIDADTRAMOANT	IDFESTACION
4	336284	10b	1	7	I	2061.134509	37.18	4_278

Tabla 5.34: Registro de muestra del conjunto de datos M4 TRAMOS.

- **M10 TRAMOS**: representa los tramos de rutas de Metro Ligero ofrecidas en la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.35.

MODO	CODITINERARIO	NUMLINEAUSUARIO	SENTIDO	NUMORDEN	TIPOPARADA	LONGITUDTRAMOANT	VELOCIDADTRAMOANT	IDFESTACION
10	335994	3	2	10	I	510.4310	25.0	10_29

Tabla 5.35: Registro de muestra del conjunto de datos M10 TRAMOS.

- **M5 TRAMOS**: representa los tramos de rutas de Cercanías ofrecidas en la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.36.

MODO	CODITINERARIO	NUMLINEAUSUARIO	SENTIDO	NUMORDEN	TIPOPARADA	LONGITUDTRAMOANT	VELOCIDADTRAMOANT	CODESTACION
5	330357	C-2	2	14	I	4495.2335	30.0	106

Tabla 5.36: Registro de muestra del conjunto de datos M5 TRAMOS.

- **M6 TRAMOS**: representa los tramos de rutas de Autobuses EMT ofrecidas en la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.37.

MODO	CODITINERARIO	NUMLINEAUSUARIO	SENTIDO	NUMORDEN	TIPOPARADA	LONGITUDTRAMOANT	VELOCIDADTRAMOANT	CODESTACION
6	283679	139	1	18	I	379.1970	30.0	2802

Tabla 5.37: Registro de muestra del conjunto de datos M6 TRAMOS.

- **M8 TRAMOS**: representa los tramos de rutas de Autobuses Interurbanos ofrecidas en la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.38.

MODO	CODITINERARIO	NUMLINEAUSUARIO	SENTIDO	NUMORDEN	TIPOPARADA	LONGITUDTRAMOANT	VELOCIDADTRAMOANT	CODESTACION
8	353765	313	1	5	I	5667.2127	40.0	17491

Tabla 5.38: Registro de muestra del conjunto de datos M8 TRAMOS.

Conjunto de Datos de Viajes (ENERO VIAJES 2019). Representa el fichero de las transacciones de viaje realizadas en Enero de 2019 por tarjetas de transporte asociadas a pasajeros de la red de transporte público de Madrid. Un registro de muestra de este conjunto aparece en la Tabla 5.39.

TARJETA	FECHA	TUSUARIO	TITULO	DESCUENTO	DPAYPOINT	IDTLV	CODVAL
000009243C6222BEF399E547EEFCA81C	14/01/2019 07:25:20	03	1055	00	02_L2_P54	C0	1

Tabla 5.39: Registro de muestra del conjunto de datos ENERO VIAJES 2019.

Nota: el identificador de la tarjeta de transporte ha sido anonimizado mediante la aplicación de una función hash.

5.2. Transformación

En esta parte del proceso ETL se implementan todas las operaciones de transformación planificadas en el *logical datamap* diseñado durante la etapa de Análisis del proyecto.

La realización de estas operaciones de tranformación tiene como objetivo preparar los datos iniciales para convertirlos en los requeridos por el proyecto. Los tipos de tareas consideradas dentro de este proceso de manipulación de datos son variadas e incluyen operaciones de limpieza, ajuste e integración de datos para asegurar la disposición de unos datos consistentes, fiables y de calidad que puedan ser utilizados para aplicar sobre ellos el método *trip chaining* con garantías.

A continuación, se describen las principales transformaciones realizadas sobre los diferentes grupos de conjuntos de datos iniciales.

Conjuntos de Datos de Estaciones.

- **Renombrar atributos de las estaciones**: consiste en cambiar el nombre de los atributos de los conjuntos de datos de estaciones para su unificación.
- **Unión de conjuntos de estaciones**: consiste en la concatenación de los conjuntos de datos en un mismo conjunto de estaciones.
- **Integración de estaciones con datos de topología**: consiste en la integración con los datos de topología, a través del identificador de estación, para obtener las coordenadas geográficas (latitud/longitud) asociadas a las estaciones. Estas coordenadas están descritas en el sistema de coordenadas WSG84 [49], correspondiente al estándar GPS y definido por el código de sistema de referencia EPSG:4326. Más información acerca de los sistemas de referencia geográficos se puede consultar en la bibliografía [50].
- **Imputación de valores de LATITUD y LONGITUD de estaciones**: representa la imputación de las coordenadas de latitud y longitud asociadas a los registros de estaciones que no han podido ser integrados con algún registro de topología. Se siguen 2 posibles aproximaciones para hallar sus valores. Por un lado, se intenta convertir las coordenadas expresadas por los atributos X e Y, descritas en el sistema de coordenadas ED50 / UTM zone 30N con código de sistema de referencia EPSG:23030, en coordenadas GPS. Por otro lado, si los atributos X e Y no tienen un valor, se recurre a la asignación manual de las coordenadas geográficas de las estaciones.

- **Cálculo de la distancia entre estaciones**: consiste en el cálculo de la distancia ortodrómica[2] existente entre cada par de estaciones de la red de transporte público de Madrid.

Conjuntos de Datos de Topologías.

- **Renombrar atributos de topologías**: consiste en cambiar el nombre de los atributos de los conjuntos de datos de topologías para su unificación.
- **Unión de conjuntos de topologías**: consiste en la concatenación de los conjuntos de topologías en un mismo conjunto de datos.
- **Eliminación de registros duplicados de topologías**: consiste en la eliminación de los registros duplicados existentes antes de llevar a cabo la integración con los datos de estaciones.

Conjuntos de Datos de Tramos de Rutas.

- **Renombrar atributos de tramos de rutas**: consiste en cambiar el nombre de los atributos de los conjuntos de datos de tramos de rutas para su unificación.
- **Unión de conjuntos de tramos de rutas**: consiste en la concatenación de los conjuntos de tramos de rutas en un mismo conjunto de datos.
- **Imputación de valores y resolución de inconsistencias en VELOCIDADTRAMOANTERIOR de tramos de rutas**: consiste en la imputación de los valores desconocidos y la corrección de las inconsistencias del atributo VELOCIDADTRAMOANTERIOR. La estrategia a seguir en ambos casos consiste en sustituir el valor de velocidad incorrecto por la media de velocidad en el modo de transporte considerado en cada caso.
- **Creación del atributo TIEMPOTRAMOANTERIOR de tramos de rutas**: consiste en el creación del atributo TIEMPOTRAMOANTERIOR a partir de los valores correspondientes de los atributos LONGITUDTRAMOANTERIOR y VELOCIDADTRAMOANTERIOR. Este nuevo campo contendrá el valor de tiempo, en minutos, asociado a cada tramo de ruta.
- **Creación del conjunto de rutas**: consiste en la creación del conjunto de datos asociado a las rutas ofrecidas en la red de transporte público de Madrid.
- **Creación del conjunto de paradas de rutas**: consiste en la creación del conjunto de datos asociado a las paradas de ruta por las que transcurren las rutas ofrecidas en la red de transporte público de Madrid.

[2] https://es.wikipedia.org/wiki/Ortodromica

- **Creación del atributo ID de paradas de rutas**: consiste en la creación del identificador de las paradas de ruta a partir de la concatenación de los atributos CODIGOITINERARIO e IDESTACION.
- **Creación del atributo DPAYPOINT de paradas de rutas**: consiste en la obtención del dpaypoint asociado a cada parada de ruta consultando el conjunto de datos de topologías. Según el modo de transporte al que pertenezca la parada de ruta, la estrategia para la derivación del dpaypoint varía.
- **Imputación de valores de DPAYPOINT de paradas de rutas**: consiste en la imputación manual de los dpaypoint de las paradas de ruta para las que no se había podido obtener previamente.
- **Resolución de inconsistencias en TIPOPARADA de paradas de rutas**: consiste en la resolución manual de los valores incorrectos en el atributo TIPOPARADA de ciertas paradas de ruta, para poder calcular posteriormente el tiempo de trayecto entre los pares de paradas.
- **Creación del conjunto de tiempos entre paradas de rutas**: consiste en la creación del conjunto de datos asociado a los tiempos entre paradas de ruta, para ello se calcula el tiempo de trayecto entre cada par de paradas consecutivas en una misma ruta.

Conjunto de Datos de Viajes (ENERO VIAJES 2019).

- **Creación del atributo DIAVIRTUAL de transacciones de viaje**: consiste en la creación del atributo DIAVIRTUAL de cada transacción de viaje teniendo en cuenta un rango horario diferente al habitual (de 5:00h a 4:59h).
- **Eliminación de las transacciones de viaje asociadas a tarjetas de empleados**: consiste en la eliminación de todas las transacciones asociadas a las tarjetas de transporte de empleados de la red de transporte público de Madrid, pues se considera que no representan los patrones de viaje de los pasajeros comunes.
- **Eliminación de las transacciones de viaje asociadas a tarjetas de excepciones**: consiste en la eliminación de todas las transacciones asociadas a las tarjetas de excepciones, es decir, tarjetas cuya frecuencia de transacciones supera el umbral de 20 transacciones de viaje en un mismo día virtual, pues se considera que no representan los patrones de viaje de los pasajeros comunes. Una posible explicación a este tipo de tarjetas puede ser que su identificador de tarjeta sea genérico, es decir, empleado en más de una tarjeta de transporte.
- **Ordenación del conjunto de transacciones de viaje por DIAVIRTUAL**: consiste en la ordenación del conjunto completo de transacciones de viaje por día virtual.

- **Eliminación de las transacciones de viaje con inconsistencias**: consiste en la eliminación de las transacciones que presenta inconsistencias en alguno de sus atributos. En este caso, las causas de inconsistencia consideradas son debidas a la existencia de transacciones únicas en un día, de códigos de validación erróneos o no reconocidos, de dpaypoints no reconocidos o transacciones coincidentes en día virtual y tarjeta con otras descartadas por alguna de las causas anteriores.
- **Ordenación del conjunto de transacciones de viaje por TARJETA**: consiste en la ordenación del conjunto completo de transacciones de viaje por tarjeta de transporte.
- **Creación de los atributos TIPO y MODO de transacciones de viaje**: consiste en la creación de los atributos TIPO y MODO de las transacciones de viaje a través de su deducción a partir de los atributos DPAYPOINT y CODVAL, respectivamente.
- **Eliminación de los atributos IDTLV y CODVAL de transacciones de viaje**: eliminación de los atributos IDTLV y CODVAL de las transacciones de viaje con el objetivo de reducir el espacio ocupado por el conjunto completo de transacciones.
- **Creación y preprocesamiento del conjunto reducido de transacciones de viaje de los días 14/01/2019 y 15/01/2019**: consiste en la creación y preprocesamiento de un conjunto reducido de las transacciones de viaje asociadas a los días virtuales 14/01/2019 y 15/01/2019 para su utilización en la aplicación del método *trip chaining*.

Esta fase de transformación de datos, así como el proceso ETL completo, se lleva a cabo tanto sobre el conjunto global de transacciones de viaje correspondientes al mes de enero de 2019 como sobre un conjunto reducido de dichas transacciones correspondientes a los días 14 y 15 de enero de 2019, el cual se empleará para implementar el método *trip chaining* sobre él. Esta decisión de reducir el conjunto de datos de transacciones se debe a que la ejecución del método *trip chaining* conlleva un gran coste computacional, al necesitar recorrer de forma secuencial todas las transacciones de viaje de cada día virtual agrupadas por tarjeta de transporte de pasajero.

A continuación, se presenta la Tabla 5.40 en la que se muestra un resumen con la reducción en el número de transacciones de viaje de cada conjunto considerado al aplicar las operaciones de eliminación del proceso ETL.

5.3. Carga

Respecto a la carga de datos, los conjuntos de datos transformados son escritos en distintos ficheros de cara a su lectura posterior cuando se aplique el método *trip chaining*. El tamaño y número de registros de cada uno de estos ficheros obtenidos puede consultarse en la Tabla 5.41.

Conjunto de Datos	Nº Transacciones Inicial	Nº Transacciones tras ETL
Global - 01/01/2019 a 31/01/2019	**151.094.796**	**109.101.600** (72 % del total)
Reducido - 14/01/2019 a 15/01/2019	**11.915.318**	**8.791.867** (74 % del total)

Tabla 5.40: Estadísticas Resumen del número de transacciones resultantes del proceso ETL.

Fichero de datos	Nº de registros	Tamaño
estaciones_df.csv	13.672	847 KB
rutas_df.csv	43.140	977 KB
paradas_rutas_df.csv	54.135	7,43 MB
tiempo_entre_paradas_rutas_df.csv	54.134	2,87 MB
viajes_preprocesado_etl.txt	109.101.600	8,65 GB

Tabla 5.41: Tamaño y número de registros de cada fichero de datos resultante del proceso ETL.

A continuación, se muestra un registro por fichero de datos con el fin de ilustrar la estructura de los datos obtenidos de la ejecución del proceso ETL. Si bien el propósito es que todos los conjuntos de datos resultantes del proceso sigan la estructura diseñada en la etapa de Análisis del proyecto, hay que tener en cuenta que el conjunto de transacciones de viaje aún debe ser procesado a través del método *trip chaining*.

- **estaciones_df.csv**: contiene el conjunto de estaciones de la red de transporte público de Madrid siguiendo la estructura diseñada para la tabla ESTACIÓN. Un registro de muestra de este fichero de datos aparece en la Tabla 5.42.

ID	NOMBRE	MUNICIPIO	DISTRITO	CORONA	LATITUD	LONGITUD
4_38	NOVICIADO	079	01	A	40.42484	-3.70742

Tabla 5.42: Registro de muestra del fichero estaciones_df.csv.

- **rutas_df.csv**: contiene el conjunto de rutas ofrecidas en la red de transporte público de Madrid siguiendo la estructura diseñada para la tabla RUTA. Un registro de muestra de este fichero de datos aparece en la Tabla 5.43.

MODO	ID	NUMERO_LINEA	SENTIDO	NUMERO_SUBLINEA
4	336284	10b	1	

Tabla 5.43: Registro de muestra del fichero rutas_df.csv.

- **paradas_rutas_df.csv**: contiene el conjunto de paradas por las que transcurren las rutas ofrecidas en la red de transporte público de Madrid siguiendo la estructura diseñada para la tabla PARADA_RUTA. Un registro de muestra de este fichero de datos aparece en la Tabla 5.44.

ID_RUTA	ORDEN	TIPO	ID_ESTACION	ID	DPAYPOINT
336284	7	I	4_278	336284:4_278	02_L10_P55

Tabla 5.44: Registro de muestra del fichero paradas_rutas_df.csv.

- **tiempo_entre_paradas_rutas_df.csv**: contiene el conjunto de tiempos entre paradas de ruta de la red de transporte público de Madrid siguiendo la estructura diseñada para la tabla TARDAR. Un registro de muestra de este fichero de datos aparece en la Tabla 5.45.

ID_PARADA_RUTA_1	ID_PARADA_RUTA_2	ID_RUTA	TIEMPO	ORDEN
336284:4_284	336284:4_283	336284	2.5116084756958093	1

Tabla 5.45: Registro de muestra del fichero tiempo_entre_paradas_rutas_df.csv.

- **viajes_preprocesado_etl.txt**: contiene el conjunto de transacciones de viaje preprocesado para su uso en la aplicación del método *trip chaining*. Un registro de muestra de este fichero de datos aparece en la Tabla 5.46.

TARJETA	FECHA	DPAYPOINT	DIAVIRTUAL	MODO	TIPO
000009243C6222BEF399E547EEFCA81C	14/01/2019 07:25:20	02_L2_P54	14/01/2019	4	SUBIDA

Tabla 5.46: Registro de muestra del fichero viajes_preprocesado_etl.txt.

Implementación del método *trip chaining*

A lo largo de este capítulo, se abordará el proceso completo seguido para la aplicación del método *trip chaining* desde los datos refinados obtenidos tras la ejecución del proceso ETL hasta la obtención de la información de etapas y viajes reconstruidos para cada tarjeta de transporte considerada, los cuales consolidarán el *smart data* final del proyecto a evaluar. La adaptación de la aplicación del método *trip chaining* sobre la red de transporte público de Madrid tendrá en cuenta ciertos valores de parámetros que pueden parecer arbitrarios, pero que han sido interpretados a partir de estadísticas de movilidad proporcionadas por Moovit, una aplicación web reconocida en temas de movilidad urbana [3]. En cuanto a la implementación completa del método *trip chaining*, en este proyecto se utiliza el lenguaje de programación Python por medio de un Jupyter Notebook.

6.1. Consideraciones iniciales

La propuesta sugerida en este proyecto para reconstruir los flujos de viaje de los pasajeros de la red de transporte público de Madrid y, en consecuencia, poder elaborar las matrices OD de tránsito generales es la aplicación del método estadístico *trip chaining*. Este método permite estimar las paradas de bajada de los segmentos de viaje a partir de las transacciones de subida registradas, las cuales es factible obtener a través de los terminales AFC instalados a lo largo de la red de transporte. Este es el escenario en el se encuadran los datos de transacciones de viajes en la red de transporte de Madrid proporcionados por el CRTM.

En este proyecto, se llevará a cabo esta estimación adaptando las suposiciones básicas del método *trip chaining* y considerando la aplicación de una serie de reglas para la detección de los transbordos entre segmentos de viaje. La detección correcta de los transbordos que

[3] https://moovitapp.com/insights/es/Moovit_Insights_Índice_de_Transporte_Público_Espana_Madrid-21

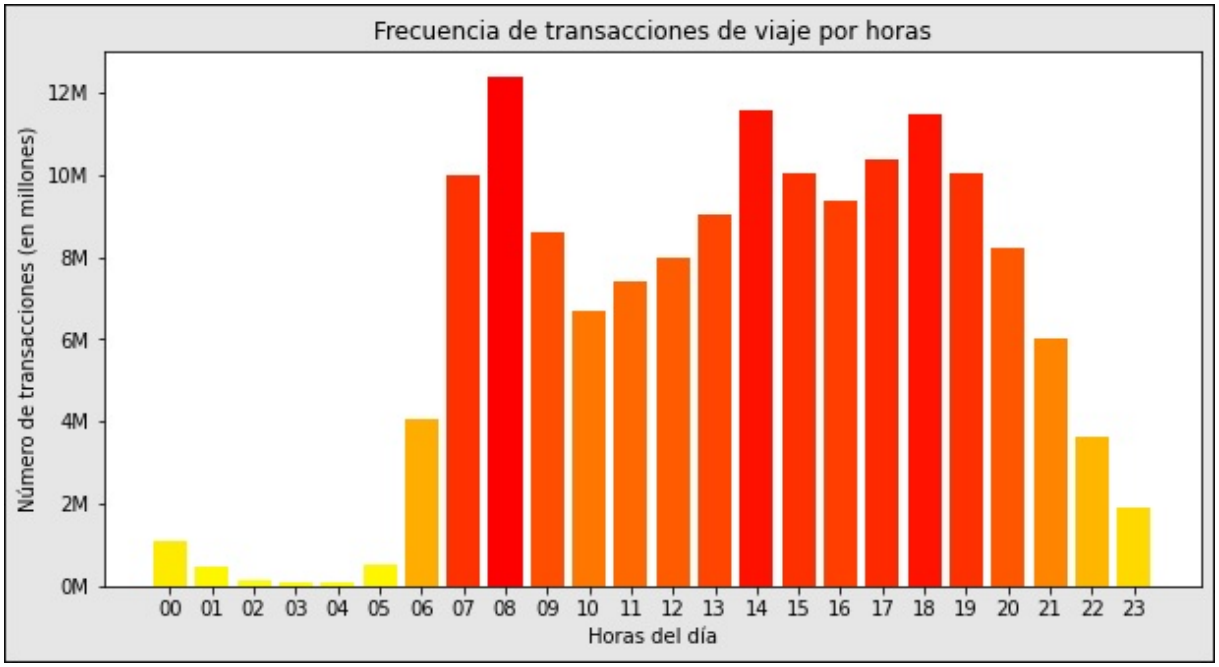

Figura 6.25: Frecuencia de transacciones de viaje por horas.

realiza un pasajero en la red de transporte es un aspecto crucial para reconstruir sus viajes origen-destino adecuados, pues una incorrecta consideración de los mismos conllevaría resultados espurios en las matrices OD de tránsito generales.

Una adaptación importante dentro de la solución propuesta consiste en considerar días virtuales con rango horario diferente al estándar establecido de comenzar los días a las 00:00h y terminarlos a las 23:59h. Así, un día virtual daría comienzo a las 05:00h y terminaría a las 04:59h del siguiente día natural, con el objetivo de adaptar un rango horario más realista en el que los pasajeros realizan sus viajes. La decisión de adoptar este nuevo rango horario radica en el análisis realizado de la frecuencia de transacciones de viaje registradas por hora en los terminales AFC de la red de transporte de Madrid. En la Figura 6.25 se presenta un histograma elaborado para mostrar la frecuencia de transacciones de viaje registradas por hora durante el período de estudio. A partir de esta figura se deduce que la franja horaria de las 04:00h a las 05:00h se corresponde aproximadamente con la de la terminación de los viajes del día en curso, al ser la franja con menos frecuencia de transacciones de viaje.

6.2. Prepreprocesamiento de las transacciones de viaje condicionado por *trip chaining*

El fichero de transacciones de viaje derivado del proceso ETL presenta aún ciertas deficiencias que deben ser resueltas previamente a la ejecución del método *trip chaining*. La

corrección de dichas transacciones de viaje requiere el recorrido secuencial de todas ellas, lo cual impone una restricción de coste computacional importante, al estar desarrollando el proyecto con los recursos limitados de una máquina virtual. El gran volumen de transacciones de viaje recopiladas (del orden de 100 millones de registros) motiva la necesidad de llevar a cabo el preprocesamiento solamente sobre un subconjunto de todas las transacciones de viaje registradas en el mes de enero de 2019. En este caso, el subconjunto seleccionado se corresponde con las transacciones de viaje registradas únicamente los días 14 y 15 de enero de 2019.

Las principales operaciones realizadas como parte del preprocesamiento de estas transacciones de viaje se resumen a continuación:

- **Eliminación de transacciones duplicadas**: consiste en descartar transacciones asociadas a una tarjeta de transporte en el día virtual considerado que se produzcan en la misma estación, se registren prácticamente al mismo tiempo (margen de tiempo de 1 minuto) y cuyo modo de transporte asociado sea el mismo. Al disponer las diferentes paradas dentro de una misma estación de distintos identificadores y coordenadas geográficas, la consideración de que 2 transacciones se produzcan en una misma estación se resuelve haciendo un cálculo de la distancia existente entre las paradas de las transacciones en cuestión. Así, he decidido determinar que se trata de una misma estación cuando la distancia entre las paradas de ambas transacciones es menor que 200 metros. El objetivo del establecimiento de este margen de distancia es asegurarse de obtener todas las paradas que pertenecen a la misma estación, con el objetivo de determinar si las transacciones registradas al mismo tiempo están duplicadas. En caso contrario, si se considera que las paradas son de distinta estación, habría que proceder a descartar el conjunto de las transacciones asociadas a la tarjeta de transporte por contener transacciones de viaje que se producen al mismo tiempo en distintas estaciones.
- **Eliminación de las transacciones únicas en un día de una tarjeta**: consiste en descartar las transacciones únicas asociadas a una tarjeta de transporte en el día virtual considerado. Esta decisión se debe a la necesidad por parte del método *trip chaining* de disponer de más de 1 transacción para poder reconstruir los viajes realizados con la tarjeta de transporte.
- **Eliminación de transacciones de una tarjeta que sean anteriores a la primera transacción de subida**: consiste en descartar aquellas transacciones de viaje asociadas a una tarjeta de transporte en el día virtual considerado que se registren antes de la primera transacción de subida. Esta decisión es motivada porque *trip chaining* estima la parada de bajada del último segmento de viaje del día teniendo en cuenta la primera transacción de subida. De esta forma, se evita comenzar el procesamiento de las transacciones de las tarjetas a partir una transacción que no sea subida y que, por tanto, pertenezca a un viaje anterior que no podrá ser completado.

- **Reordenación de transacciones de una tarjeta**: consiste en cambiar el orden de ciertos pares consecutivos de transacciones asociadas a una tarjeta de transporte en el día virtual considerado si sus tipos llevan a deducir que el orden lógico es el contrario. Esta situación se presenta principalmente en los pares de transacciones donde la de bajada aparece inmediatamente después de la de subida o transbordo.
- **Eliminación de transacciones inconsistentes**: consiste en descartar todas las transacciones asociadas a una tarjeta de transporte en el día virtual considerado que presentan inconsistencias en su estructura que hacen inviable la aplicación del método *trip chaining* sobre ellas. Cuando aparecen este tipo de conjuntos de transacciones anómalas, la decisión que se toma es invalidar todas ellas. A nivel de implementación, dentro de esta operación general de eliminación de transacciones se engloban las siguientes:
  - **Eliminación de todas las transacciones de una tarjeta cuando hay menos de 2 subidas**: consiste en descartar todas las transacciones asociadas a una tarjeta de transporte en el día virtual considerado en caso de no contener un mínimo de 2 transacciones de subida. Esta decisión se debe a la imposición por parte de *trip chaining* de tener un mínimo de 2 transacciones de subida para poder reconstruir los viajes de un pasajero.
  - **Eliminación de todas las transacciones de una tarjeta cuando algún par consecutivo de transacciones presenta incoherencias en su registro**: consiste en descartar todas las transacciones asociadas a una tarjeta de transporte en el día virtual considerado en caso de encontrar algún par de transacciones consecutivas que tengan lugar prácticamente al mismo tiempo (margen de tiempo de 1 minuto) y se produzcan en ubicaciones distantes más de 200 metros, posiblemente correspondientes a estaciones distintas.
  - **Eliminación de todas las transacciones de una tarjeta cuando algún par consecutivo de transacciones no es reordenable**: consiste en descartar todas las transacciones de viaje asociadas a una tarjeta de transporte en el día virtual considerado en caso de existir algún par consecutivo de transacciones que no siga un orden lógico de registro y, además, no pueda ser reordenado de forma coherente.

Una vez llevado a cabo el preprocesamiento para resolver las anomalías presentes en las transacciones, el fichero de transacciones de viaje resultante ya se encuentra en un estado óptimo para su utilización en la aplicación del método *trip chaining*.

Esta fase de preprocesamiento de transacciones ha provocado una reducción del número de transacciones de viaje correspondientes a los días 14 y 15 de enero de 2019 disponibles para ejecutar sobre ellas el método *trip chaining*. A continuación, en la Tabla 6.47 se muestra un resumen del número de transacciones resultantes del preprocesamiento junto con la frecuencia de cada una de las causas de eliminación de dichas transacciones. Igualmente,

Nº Transacciones Inicial	**8.791.867**
Nº Transacciones eliminadas	**246.468**
- Transacciones duplicadas	74.219
- Transacciones únicas en un día	12.969
- Transacciones inconsistentes	139.502
- Relacionadas con las anteriores (tarjeta y día coincidentes)	19.778
Nº Transacciones Final	**8.545.389** (97 % del total)

Tabla 6.47: Estadísticas Resumen con las causas de descarte de transacciones de viaje debidas al preprocesamiento condicionado por *trip chaining*.

una explicación del significado de estas causas de eliminación de transacciones de viaje puede consultarse en la Tabla 6.48.

Causa Eliminación de Transacciones	**Descripción**
Transacciones duplicadas	Correspondiente a transacciones que se produzcan en la misma estación, se registren prácticamente al mismo tiempo (margen de tiempo de 1 minuto) y cuyo modo de transporte asociado sea el mismo.
Transacciones únicas en un día	Correspondiente a tarjetas que solo tienen 1 transacción no duplicada en el día considerado.
Transacciones inconsistentes	Correspondiente a conjuntos de transacciones de tarjeta cuya estructura hace inviable la aplicación del método *trip chaining* sobre ellas.
Relacionadas con las anteriores	Correspondiente a transacciones cuya tarjeta y día asociados son iguales que los de alguna de las transacciones eliminadas por una de las causas anteriores.

Tabla 6.48: Explicación de las causas de descarte de transacciones de viaje condicionadas por *trip chaining*.

6.3. Datos de entrada

Los datos de transacciones resultantes del preprocesamiento anterior serán los empleados como entrada del método *trip chaining* a aplicar en esta parte del proyecto.

A continuación, como ejemplo, se seleccionan las transacciones de viaje asociadas a una tarjeta de transporte y registradas durante un mismo día virtual para ilustrarlas visualmente sobre un mapa. Por un lado, en la Tabla 6.49 se listan los registros de las transacciones de viaje seleccionadas correspondientes a la actividad de una tarjeta de

TARJETA	FECHA	DPAYPOINT	DIAVIRTUAL	MODO	TIPO
000C16E77B1425A74252AD1FEFB5CFB5	14/01/2019 10:16:40	02_L12_P27	14/01/2019	4	SUBIDA
000C16E77B1425A74252AD1FEFB5CFB5	14/01/2019 10:26:06	04_L0_P433	14/01/2019	5	SUBIDA
000C16E77B1425A74252AD1FEFB5CFB5	14/01/2019 10:59:16	04_L0_P7	14/01/2019	5	BAJADA
000C16E77B1425A74252AD1FEFB5CFB5	14/01/2019 15:03:14	02_L1_P16	14/01/2019	4	SUBIDA

Tabla 6.49: Transacciones de viaje de una tarjeta de transporte seleccionadas como ejemplo.

transporte durante el día virtual de viaje 14/01/2019. Por otro lado, en la Figura 6.26 se representa la misma información de manera más visual, sobre un mapa, con el objetivo de ubicar las transacciones en las estaciones donde se producen. En esta figura se puede observar la presencia en el mapa de los 4 marcadores correspondientes a cada una de las transacciones de viaje seleccionadas. La ubicación de estos marcadores es obtenida a través de la parada asociada al *dpaypoint* de cada transacción. Además, en el margen de la figura, se puede comprobar en la leyenda el tipo de cada transacción representada.

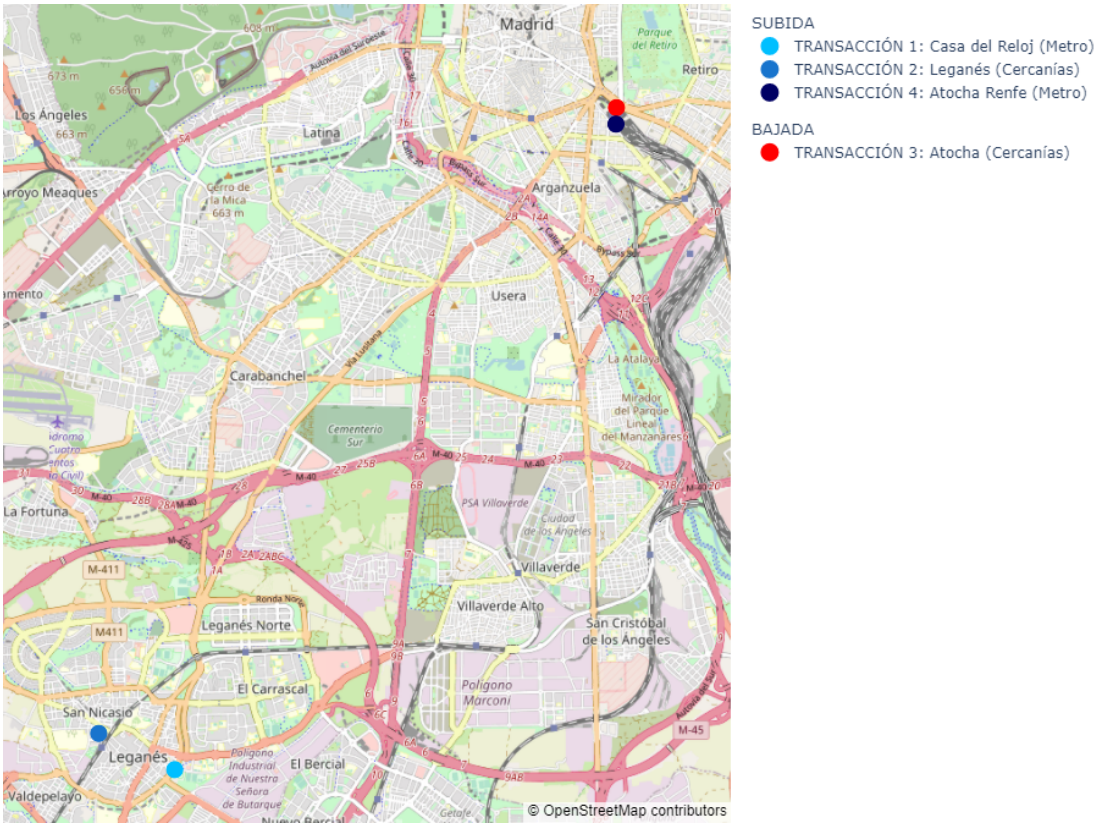

Figura 6.26: Representación sobre un mapa de las transacciones de viaje de una tarjeta seleccionadas como ejemplo.

Posteriormente, en la sección 6.7, se mostrarán las etapas y viajes reconstruidos que se obtienen al aplicar el método *trip chaining* sobre estas mismas transacciones de viaje.

6.4. Estimación de la parada de bajada

La estimación de la parada de bajada correspondiente a cada segmento de viaje recorrido por los pasajeros de una red de transporte es el problema que el método *trip chaining* está destinado a resolver. Afrontar la resolución de este problema conlleva la asunción de una serie de suposiciones que deben ser adaptadas para su aplicación sobre la red de transporte considerada. En este caso, las suposiciones adaptadas del método *trip chaining* para su aplicación sobre la red de transporte público de Madrid se enuncian a continuación:

- **Suposición de continuidad adaptada**: la adaptación de esta suposición del método consiste en considerar como posibles paradas de bajada de un segmento de viaje aquellas que se encuentren dentro de un radio de distancia umbral máxima alrededor de la siguiente parada de subida y, además, que sean del mismo de modo transporte en el que se inició el segmento de viaje. Esta decisión se toma considerando que la siguiente subida de un pasajero en el transporte público queda registrada por una nueva transacción asociada a su tarjeta de transporte.
- **Suposición de simetría adaptada**: la adaptación de esta suposición del método consiste en considerar días virtuales de viaje con rango horario diferente al habitual (de 5:00h a 4:59h) con el objetivo de obtener una primera subida del día del pasajero más realista acorde con el estudio de la frecuencia general de transacciones de viaje de pasajeros mostrada en la sección anterior.

Contando con estas suposiciones de partida, el método *trip chaining* pretende dar una estimación de la parada de bajada de un segmento de viaje teniendo en cuenta la información de la siguiente transacción de subida o, en caso de tratarse del último segmento de viaje, de la primera transacción de subida del día. La resolución de este problema es considerada como un proceso de búsqueda de las posibles paradas en la red de transporte que pueden ser alcanzadas desde la parada de subida del segmento de viaje, teniendo en cuenta tanto la ruta de transporte seguida como el modo de transporte en el que se realiza. Así, el proceso seguido para la determinación de las paradas de bajada adaptado para la red de transporte público de Madrid puede ser resumido en 3 fases:

Obtención de las paradas más cercanas. En primer lugar, se procede a la obtención de las paradas en la red de transporte que se encuentran dentro de un radio de distancia umbral máxima alrededor de la parada de subida siguiente, en lo que comúnmente es denominado en la literatura como *Buffer Zone*. En este proyecto, se considera una distancia umbral máxima de 650 metros, determinada por la distancia media que los pasajeros de la red de transporte de Madrid caminan por viaje para llegar a su parada de subida. La

decisión de adoptar este valor concreto no es fortuita, sino que se deriva de estadísticas de movilidad en el transporte público aportadas por Moovit.

Como resultado de esta fase se obtiene el conjunto de todas las paradas de la red de transporte que se encuentran dentro de una distancia de 650 metros desde la parada de subida siguiente.

Filtrado de las paradas alcanzables desde la parada de subida. Seguidamente, el conjunto de paradas más cercanas debe ser reducido a únicamente las que sean alcanzables desde la parada de subida asociada al segmento de viaje actual. Un primer filtrado de estas paradas es relativo al modo de transporte en el que se haya registrado la transacción de subida asociada al segmento de viaje actual. Ello implica una limitación del conjunto de posibles paradas de bajada a únicamente las correspondientes al modo de transporte considerado.

Por otro lado, se plantea un segundo filtrado del conjunto de paradas, con el objetivo de poder estimar la parada de bajada correcta. A este respecto, se consideran 2 enfoques diferenciados cuya aplicación dependerá del modo de transporte en el que se lleve a cabo el segmento de viaje cuya parada de bajada debe ser estimada.

- **Filtrar por las paradas de la ruta actual**: este enfoque se emplea en los modos de transporte de metro ligero, autobuses EMT y autobuses interurbanos, donde la información de la ruta seguida es visible y la de transbordos quedará registrada a través de nuevas transacciones de subida en el transporte público. De esta manera, se propone filtrar el conjunto de paradas por las que transcurran a lo largo de la misma ruta de transporte seguida desde la parada de subida del segmento de viaje actual. Siguiendo este enfoque se consigue limitar el conjunto inicial a las paradas de una única ruta.
- **Determinar todas las paradas como alcanzables**: este enfoque es utilizado cuando el modo de transporte considerado es el metro o cercanías, en los cuales no se dispone de información alguna acerca de la ruta seguida al entrar en la estación de subida. Igualmente, la información de transbordos en estos modos se considera transparente, pues los pasajeros pueden moverse con libertad dentro de sus instalaciones y sin restricciones entre sus rutas disponibles. A nivel de implementación, se consideran todas las rutas de metro y las de cercanías como rutas únicas, por las que los pasajeros pueden moverse libremente. Siguiendo este enfoque el conjunto de posibles paradas no es reducido, pues se considera que todas las paradas del modo considerado pueden ser alcanzadas desde la parada de subida inicial.

Determinación de la parada de mínima distancia. Finalmente, se selecciona como estimación de parada de bajada aquella, dentro del conjunto de paradas que ha sido reducido, que minimice la distancia con la parada de subida siguiente. De esta forma, se obtiene una estimación de la bajada que, posteriormente, podrá ser validada en los

casos donde la correspondiente transacción de bajada real esté disponible. El proceso de validación de la estimación realizada forma parte de la evaluación del proyecto y será explicado en la Sección 7.1.

6.5. Detección de transbordos

Una vez estimada la parada de bajada de un segmento de viaje, la detección de transbordos es empleada para determinar si esta parada constituye realmente la bajada de destino del viaje del pasajero o si, por el contrario, se trata de una parada intermedia de transbordo que realiza de cara a continuar el viaje actual hacia su destino.

La tarea de detección de transbordos es abordada a través de la definición de una serie de reglas cuyo cumplimiento o no determinará si se ha producido un transbordo o una actividad. Dichas reglas serán evaluadas una a una con el fin de comprobar si se satisfacen o no para determinar si efectivamente se produce un transbordo o si, por el contrario, se debe considerar la realización de una actividad por parte del pasajero y asignar el destino del viaje iniciado.

En este proyecto, se utilizan 3 reglas básicas de detección de transbordos parametrizadas específicamente para su aplicación sobre la red de transporte público de Madrid. Hay que tener en cuenta que el no cumplimiento de cualquiera de estas reglas dará lugar a la consideración de una actividad por parte del pasajero y, por tanto, significará que la parada estimada es el final de su viaje.

Regla espacial. Esta primera regla representa la distancia máxima de transbordo que puede existir entre la parada de bajada estimada y la siguiente de subida para considerar que ha podido producirse un transbordo. El parámetro que la caracteriza es referido como MTD (*maximum transfer distance*) y es establecido en 650 metros. Este valor se deriva de las estadísticas proporcionadas por Moovit con respecto a la distancia media por viaje que los pasajeros están dispuestos a andar para llegar a su parada de subida. En este caso, su valor es igual que el establecido para la distancia umbral máxima utilizada en la estimación de las paradas de bajada, pues se considera que la distancia que está dispuesto a realizar andando un pasajero para llegar tanto a su siguiente parada de subida como a la de transbordo es la misma.

Regla temporal. Esta regla representa el tiempo máximo de transbordo que puede existir entre el tiempo de bajada del segmento de viaje actual y el de subida del siguiente. El parámetro que la caracteriza es referido como MTT (*maximum transfer time*), siendo establecido en 18 minutos. La determinación de este valor concreto proviene de la suma de 2 factores distintos. Por un lado, se establece un tiempo medio de espera de 10 minutos, según lo dispuesto en las estadísticas aportadas por Moovit. Por otro lado, se realiza el cálculo del tiempo que una persona tarda en recorrer los 650 metros que, de media, está dispuesto a andar para hacer transbordo entre estaciones. Así, considerando una velocidad

media andando de 1.4 m/s [51], se obtiene un total de, aproximadamente, 8 minutos de tiempo a pie. Finalmente, la suma de ambos factores da como resultado los 18 minutos establecidos.

La evaluación de esta regla es algo más compleja al requerir calcular el tiempo de viaje transcurrido para llegar desde la parada de subida del segmento de viaje actual hasta la parada de bajada estimada. La forma de calcular el tiempo de viaje difiere dependiendo del modo de transporte involucrado en el segmento de viaje considerado. A continuación, se describe el proceso seguido para ello según las 2 diferenciaciones planteadas.

- **Cálculo del tiempo de viaje en metro y cercanías**: en el caso de los segmentos de viaje realizados en metro y cercanías, hay que tener en cuenta la posibilidad para el pasajero de moverse sin restricciones a lo largo de toda la red sin registrar transacciones relativas a los transbordos que pueda realizar entre las líneas. Este hecho abre muchas posibilidades para el pasajero de llegar desde la estación de subida hasta la correspondiente de bajada, lo cual constituye un reto a la hora de reconstruir el recorrido que sigue.

  Para resolver esta cuestión se han diseñado 2 grafos para representar la topología completa de las redes de metro y cercanías, teniendo en cuenta las estaciones de correspondencia/transbordo entre sus múltiples líneas y los tiempos de trayecto de cada uno de los tramos de ruta existentes. Así, se consigue conformar una estructura en la que la conexión entre cualquier par de estaciones dentro de las respectivas redes es posible. En este caso, la decisión tomada para determinar el trayecto seguido por un pasajero entre 2 estaciones del grafo se basa en considerar el camino más corto entre ellas, según el tiempo de viaje requerido. Para ello, se utiliza el algoritmo de Dijkstra, un popular algoritmo de Teoría de Grafos para calcular el camino más corto entre 2 nodos de un grafo. Una consideración importante a tener en cuenta es el establecimiento de un tiempo medio de espera entre estaciones a la hora de llevar a cabo un transbordo en una misma estación. En este caso, su valor asignado es de 10 minutos, acorde con lo expuesto en las estadísticas de movilidad de Moovit.

- **Cálculo del tiempo de viaje en metro ligero, autobuses EMT y autobuses interurbanos**: en el caso de los segmentos de viaje realizados en metro ligero, autobuses EMT y autobuses interurbanos, el proceso de cálculo del tiempo de viaje se simplifica, al disponer de información sobre la ruta de transporte en la que los pasajeros hacen la subida. En este caso, el tiempo de viaje transcurrido entre las paradas de subida y bajada de la ruta considerada se calcula a través de la suma de los tiempos de los tramos de viaje consecutivos que hay entre ellas.

Regla de mismo sentido de ruta. Esta última regla establece que si la paradas de bajada estimada y la siguiente de subida coinciden y siguen el mismo sentido de ruta significa que el pasajero se ha bajado del transporte público por voluntad propia, lo cual implica la llegada a su destino de viaje y anula la posibilidad de considerar un transbordo.

6.6. Ejecución de *trip chaining*.

Una vez realizado preprocesamiento de las transacciones de viaje correspondientes a los días 14 y 15 de enero de 2019, se procede a la ejecución del método *trip chaining* para reconstruir los flujos de tránsito de los pasajeros de la red de transporte público de Madrid. En este caso, la ejecución del método se ha llevado a cabo sobre una muestra de 100.000 tarjetas de transporte, de cara a obtener los resultados de la ejecución en un tiempo asumible. En concreto, han sido procesadas un total de 375.826 transacciones de viaje correspondientes a 100.000 tarjetas de transporte.

A nivel de implementación, el procesamiento del fichero de transacciones de viaje se lleva a cabo sobre particiones proporcionales del mismo, también denominadas *chunks*, a partir de las cuales se obtienen los conjuntos de transacciones que deberán ser tratados por el método *trip chaining*. Estos conjuntos de transacciones de viaje serán procesados uno a uno, recorriendo sus transacciones de forma iterativa, para aplicar sobre ellos la lógica codificada para el método *trip chaining*.

Finalmente, como resultado de la ejecución se obtienen las etapas de viaje y los viajes reconstruidos para cada tarjeta de transporte de pasajero, los cuales son escritos en ficheros de datos para su posterior evaluación para los fines del proyecto. El número de registros total en cada uno de los 2 ficheros generados junto con su tamaño puede consultarse en la Tabla 6.50.

Fichero de datos	**Nº de registros**	**Tamaño**
etapas.txt	349.550	51,1 MB
viajes.txt	257.698	32,2 MB

Tabla 6.50: Número de registros de etapas de viaje y viajes reconstruidos en los ficheros resultantes de la aplicación de *trip chaining*.

6.7. Datos de salida

Los datos de salida constituyen el resultado final del método *trip chaining* en forma de etapas de viaje y viajes reconstruidos a partir de las transacciones de viaje registradas. Esta información resultante tiene como utilidad describir los recorridos seguidos por los pasajeros de la red de transporte y determinar los viajes completos origen-destino que realizan.

A continuación, en las Tablas 6.51 y 6.52 se muestran las etapas de viaje y los viajes reconstruidos por el método *trip chaining* tras su aplicación sobre las transacciones de viaje seleccionadas como ejemplo. Asimismo, la Figura 6.27 representa dicha información sobre un mapa con el objetivo de ilustrar el recorrido seguido por el pasajero asociado a la tarjeta de transporte considerada. En esta figura se puede visualizar el recorrido seguido por el pasajero a través de las etapas de viaje reconstruidas y representadas como

líneas en el mapa. Igualmente, en la leyenda habilitada al margen de la figura se puede consultar el tipo de cada etapa de viaje. Además de la representación de las etapas de viaje reconstruidas, se muestran también sobre el mapa los viajes origen-destino resultantes de agrupar estas etapas, con el objetivo de aclarar el desplazamiento del pasajero.

ID	FECHA_HORA_INICIO	FECHA_HORA_FIN	TIPO	PARADA_RUTA_INICIO	PARADA_RUTA_FIN	ID_VIAJE
000C16E77B1425A74252AD1FEFB5CFB5: 14/01/2019:1_1	14/01/2019 10:16:40	14/01/2019 10:25:14	TRANSBORDO	4_233	4_235	000C16E77B1425A74252AD1FEFB5CFB5: 14/01/2019:1
000C16E77B1425A74252AD1FEFB5CFB5: 14/01/2019:1_2	14/01/2019 10:25:14	14/01/2019 10:26:06	TRANSBORDO	4_235	5_41	000C16E77B1425A74252AD1FEFB5CFB5: 14/01/2019:1
000C16E77B1425A74252AD1FEFB5CFB5: 14/01/2019:1_3	14/01/2019 10:26:06	14/01/2019 11:05:08	BAJADA	5_41	5_11	000C16E77B1425A74252AD1FEFB5CFB5: 14/01/2019:1
000C16E77B1425A74252AD1FEFB5CFB5: 14/01/2019:2_1	14/01/2019 15:03:14	14/01/2019 16:18:51	BAJADA	4_16	4_233	000C16E77B1425A74252AD1FEFB5CFB5: 14/01/2019:2

Tabla 6.51: Etapas de viaje reconstruidas a partir de las transacciones seleccionadas como ejemplo.

ID	FECHA_HORA_INICIO	FECHA_HORA_FIN	TIPO	TARJETA
000C16E77B1425A74252AD1FEFB5CFB5: 14/01/2019:1	14/01/2019 10:16:40	14/01/2019 11:05:08	RECONSTRUIDO	000C16E77B1425A74252AD1FEFB5CFB5
000C16E77B1425A74252AD1FEFB5CFB5: 14/01/2019:2	14/01/2019 15:03:14	14/01/2019 16:18:51	RECONSTRUIDO	000C16E77B1425A74252AD1FEFB5CFB5

Tabla 6.52: Viajes reconstruidos a partir de las transacciones seleccionadas como ejemplo.

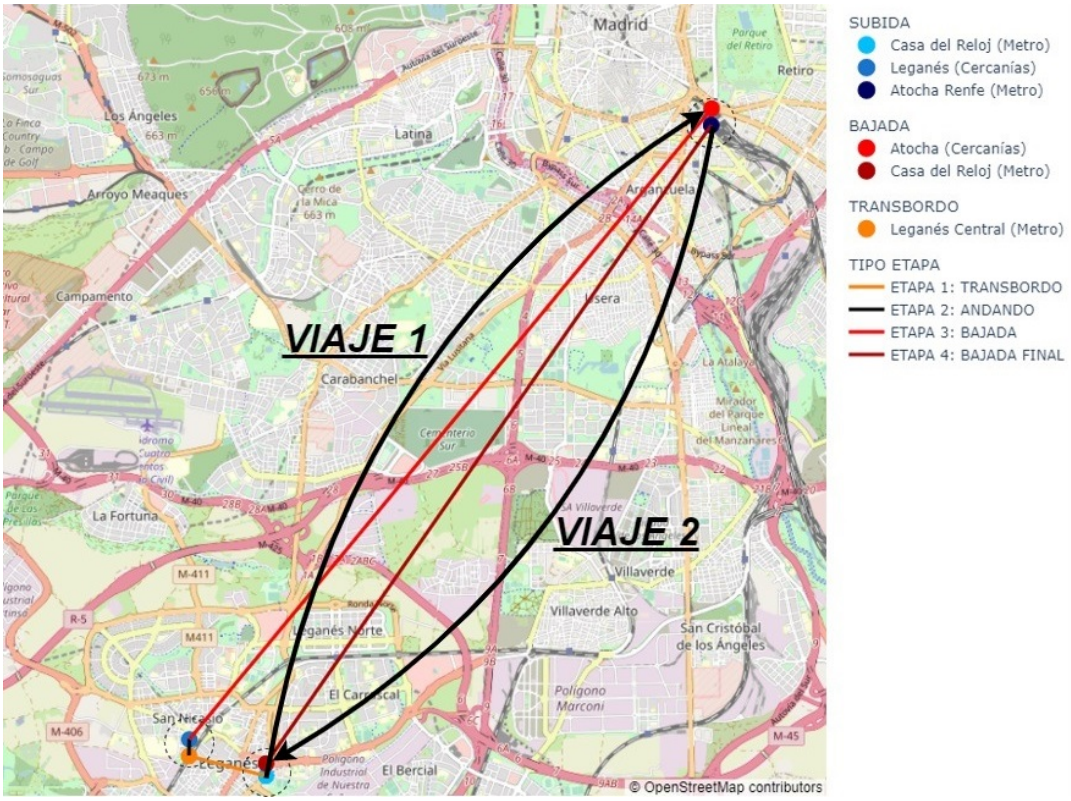

Figura 6.27: Representación sobre un mapa de las etapas de viaje y los viajes reconstruidos a partir de las transacciones seleccionadas como ejemplo.

Evaluación

En este capítulo, se llevará a cabo el proceso de evaluación y presentación de los resultados obtenidos en el proyecto. En concreto, se explicará el proceso de validación parcial seguido para valorar el funcionamiento del método *trip chaining* en su aplicación al caso particular de la red de transporte público de Madrid, junto con las estadísticas de rendimiento asociadas. Asimismo, se mostrarán los resultados más destacados que se han obtenido, tanto en formato tabular como en gráficos y mapas, con el objetivo de resumir las principales aportaciones del proyecto.

7.1. Validación

Una de las restricciones del proyecto es la imposibilidad de realizar una validación endógena completa de las estimaciones de paradas de bajada, al no disponer de todas las transacciones de bajada reales, debido a las propias características del sistema AFC de registro de transacciones implantado en la red de transporte de Madrid. A pesar de ello, sí se puede llevar a cabo una validación parcial de parte de las estimaciones, gracias a la disposición de algunas transacciones de bajada entre los datos proporcionados por el CRTM.

A la hora de validar las estimaciones de las paradas de bajada correspondientes a los segmentos de viaje reconstruidos, se considera aceptable permitir un cierto margen de error para considerar la predicción como correcta, no siendo requerido determinar la parada exacta de bajada para considerarla válida. La intuición detrás de la decisión de considerar una parada estimada como correcta, en caso de encontrarse inmediatamente antes o después de la parada de bajada real, se debe a la consideración de que puede ser habitual que los pasajeros se bajen en distintas paradas cercanas a su destino concreto, sin necesidad de ser siempre la misma. En este proyecto, se establece un margen de error de 1 parada de ruta de distancia a la hora de determinar la corrección de una estimación de bajada. No obstante, en los resultados estadísticos de validación que se mostrarán se hace la distinción entre estimaciones exactas, cuando las paradas estimada y real coincidan, y aproximadas, cuando la distancia entre ellas sea de 1 parada de ruta.

Este proceso de validación parcial planteado aporta un conjunto de estadísticas interesantes que pueden servir para tener una idea acerca del rendimiento ofrecido por el método *trip chaining* en su aplicación en la red de transporte de Madrid.

Primero, se debe tener en cuenta que el método *trip chaining* no siempre tiene la capacidad para determinar una parada de bajada en un segmento de viaje, pues pueden existir diversos impedimientos a la hora de realizar los cálculos requeridos para obtener una estimación. Los principales impedimientos encontrados son de tipo espacial y temporal. Por ejemplo, el método presenta la imposibilidad para determinar una parada de bajada que se encuentre dentro del radio de distancia umbral máxima alrededor de la parada de subida siguiente. En otras ocasiones, el método falla en la estimación cuando el tiempo requerido para completar el trayecto de viaje hasta una de las posibles paradas de bajada supera el tiempo en el que se ha registrado la parada de subida siguiente. En este proyecto, el método *trip chaining* ha podido ser aplicado con éxito para dar una estimación de la bajada en un 89,60 % de los segmentos de viaje procesados. En la Tabla 7.53 se muestra con detalle el número y porcentaje de segmentos de viaje de cada posible situación.

Tipos de Etapas de Viaje	**Número**	**Porcentaje**
Con bajada estimada	230.902	89,60 %
Inconexos	26.796	10,40 %
TOTAL	257.698	100 %

Tabla 7.53: Estadísticas de validación del método *trip chaining* por segmentos de viaje.

Por otro lado, se evalúa el rendimiento en la estimación del método *trip chaining*, con el fin determinar su tasa de acierto global, empleando las 29.536 transacciones de bajada disponibles en el conjunto de transacciones procesado durante la aplicación de *trip chaining*. En este caso, el resultado obtenido del proceso de validación parcial muestra una tasa de acierto de *trip chaining* del 81,97 %. En la Tabla 7.54 se muestra en detalle el número y porcentaje de segmentos de viaje por tipo de estimación, diferenciando entre correctas, y sus subtipos, e incorrectas.

Tipos de Estimaciones	**Número**	**Porcentaje**
Correctas	24.211	81,97 %
- Exactas	23.729	80,34 %
- Aproximadas	482	1,63 %
Incorrectas	5.325	18,03 %
TOTAL	29.536	100 %

Tabla 7.54: Estadísticas de validación del método *trip chaining* por tipo de estimación.

Además, para dar un mayor detalle acerca de los resultados obtenidos de la validación del método, se han elaborado otras tablas y mapas, a diferentes niveles de agrupación, que aportan más información de interés.

En primer lugar, los resultados de estimación han sido desglosados por los modos de transporte de cercanías, metro y metro ligero, pues son los únicos de los que se dispone de información sobre las paradas de bajada reales con las que poder validar el rendimiento del método. La Tabla 7.55 detalla los resultados de estimación relativos a cada uno de los modos anteriores, especificando el número de cada tipo de estimación junto con su porcentaje sobre el total.

Modo de Transporte	**Tipo de Estimación**	**Número de Estimaciones**	**Porcentaje**
Cercanías	Correcta	22.195	82,69 %
	- Exacta	21.837	81,36 %
	- Aproximada	358	1,33 %
	Incorrecta	4.645	17,31 %
	TOTAL	26.840	100 %
Metro	Correcta	1.470	86,52 %
	- Exacta	1.357	79,87 %
	- Aproximada	113	6,65 %
	Incorrecta	229	13,48 %
	TOTAL	1.699	100 %
Metro Ligero	Correcta	546	54,76 %
	- Exacta	535	53,66 %
	- Aproximada	11	1,10 %
	Incorrecta	451	45,24 %
	TOTAL	997	100 %

Tabla 7.55: Resultados de validación del rendimiento del método *trip chaining* por modo de transporte.

A continuación, en la Figura 7.28 se representa sobre un mapa el porcentaje de acierto del método *trip chaining* en las estaciones validadas de la red de transporte de Madrid de las que tenemos información sobre sus paradas de bajada reales. Con esta representación se puede visualizar la tasa de acierto en la estimación obtenida para cada estación y comprobar en qué estaciones el método ha obtenido mejores y peores resultados.

También, en la Figura 7.29 se representa otro mapa, a más alto nivel, que muestra la diferencia de porcentaje de acierto en la estimación, con respecto a la tasa de acierto global del método, establecida en un 81,97 %, para los distritos de Madrid Capital de los que se tiene información sobre las paradas de bajada reales. Así, podemos ver que el rendimiento del método difiere según el distrito donde se haga la estimación. Finalmente, en la Tabla 7.56 se detalla el contenido mostrado en el mapa anterior desglosando por distrito el número de estimaciones realizadas, el porcentaje de acierto y la diferencia relativa con respecto a la tasa de acierto global del método *trip chaining*.

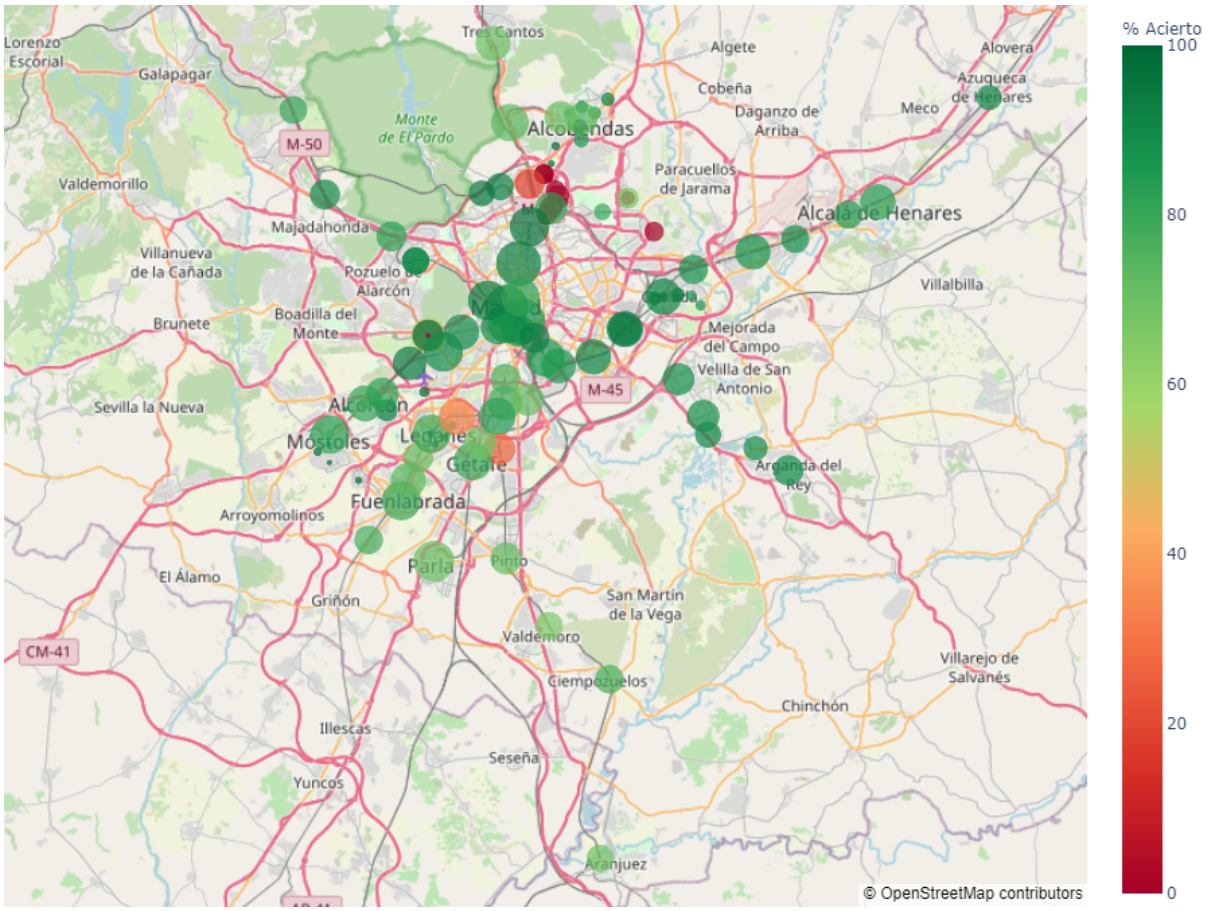

Figura 7.28: Representación sobre un mapa del porcentaje de acierto de *trip chaining* por estación validada.

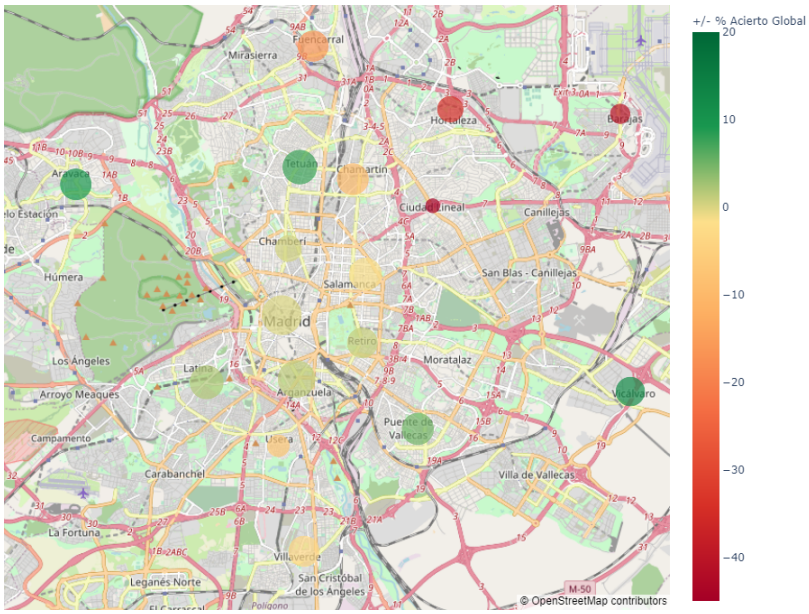

Figura 7.29: Representación sobre un mapa de la diferencia de porcentaje de acierto sobre la tasa de acierto global de *trip chaining* por distrito de Madrid Capital.

Distrito	**Número de Estimaciones**	**Porcentaje de Acierto**	**Diferencia con Tasa de Acierto Global**
Centro	8.253	82,01 %	0,04 %
Arganzuela	5.054	83,32 %	1,35 %
Retiro	1.104	82,97 %	1,00 %
Salamanca	1.762	80,59 %	-1,38 %
Chamartín	1.404	72,44 %	-9,53 %
Tetúan	2.569	89,57 %	7,60 %
Chamberí	440	82,50 %	0,53 %
Fuencarral-El Pardo	1.186	65,68 %	-16,29 %
Moncloa-Aravaca	1.296	91,44 %	9,46 %
Latina	2.228	83,98 %	2,00 %
Usera	176	75,00 %	-6,97 %
Puente de Vallecas	1.502	86,95 %	4,98 %
Ciudad Lineal	32	0,00 %	-81,97 %
Hortaleza	412	47,09 %	-34,88 %
Villaverde	1.200	78,42 %	-3,55 %
Vicálvaro	812	93,96 %	11,99 %
Barajas	100	41,51 %	-40,46 %
TOTAL	29.536	100 %	0,00 %

Tabla 7.56: Detalle de resultados de validación del rendimiento del método *trip chaining* por distrito de Madrid Capital.

7.2. Resultados del proyecto

Esta sección comprende la representación de los gráficos diseñados e implementados para mostrar los resultados más relevantes del proyecto de una forma más visual que simplemente aportando datos numéricos sobre una tabla.

La información que se va a mostrar a través de gráficos ha sido derivada a partir del fichero de viajes reconstruidos en la red de transporte público de Madrid que se generó al finalizar la ejecución del método *trip chaining*. Además, ha sido necesario realizar un proceso de integración de datos para obtener la información de diversos atributos relativa a los viajes reconstruidos, tales como el municipio, la corona tarifaria o el distrito al que se corresponden cada uno de los orígenes y destinos de dichos viajes. Finalmente, se ha llevado a cabo un proceso de agrupación y filtrado de los datos de frecuencia de viajes con el fin de elaborar unos gráficos que resuman la información más destacada según cada agrupación considerada.

7.2.1. Matrices OD de tránsito

Las matrices OD de tránsito constituyen un mecanismo de gran utilidad para visualizar de un solo vistazo la frecuencia de viajes origen-destino entre pares de ubicaciones a diferentes niveles de agrupación. Estas matrices serán el punto de partida para la realización de análisis posteriores más avanzados para propósitos más ambiciosos.

A continuación, se muestran diferentes matrices OD de tránsito de pasajeros en la red de transporte de Madrid por medio de mapas de calor, con los que representar la frecuencia de viajes origen-destino entre conexiones de pares de ubicaciones a diferentes niveles de agrupación y, también, destacar su volumen relativo respecto al conjunto de total conexiones.

Matriz OD de tránsito agrupada por corona tarifaria. La Figura 7.30 representa una matriz OD de tránsito en la que se muestra la frecuencia de viajes origen-destino que se producen entre cada par de coronas tarifarias que componen la red de transporte de Madrid. A simple vista puede comprobarse como los viajes dentro de la corona tarifaria A son los más numerosos. También se puede observar que existen ciertos pares de coronas tarifarias que presentan una conexión prácticamente residual.

Matriz OD de tránsito agrupada por municipio. La Figura 7.31 representa una matriz OD de tránsito agrupada por municipio en la que se muestran los 8 municipios con mayor volumen de viajes. En ella podemos observar como, nuevamente, los viajes dentro del municipio de Madrid son los más numerosos, destacando también, por su número, los que se producen entre la capital y los municipios de su periferia.

Matriz OD de tránsito agrupada por distrito de Madrid Capital. La Figura 7.32 se corresponde con una matriz OD de tránsito de pasajeros por distritos de Madrid

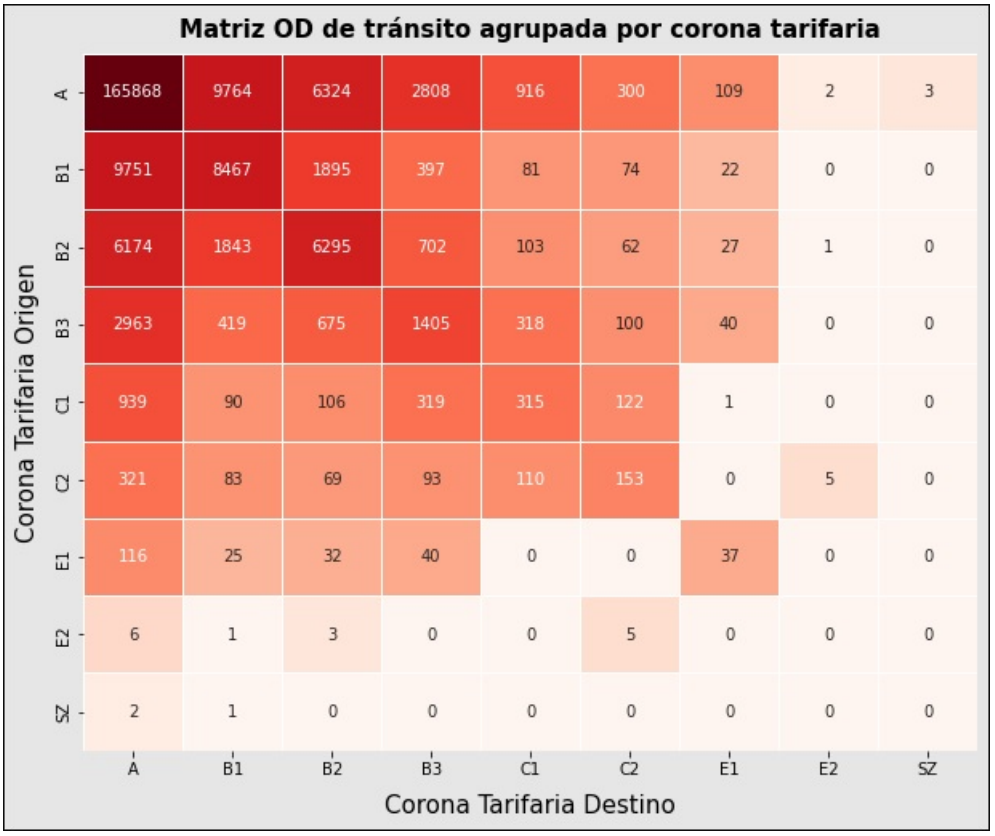

Figura 7.30: Matriz OD de tránsito agrupada por corona tarifaria.

Capital. Al igual que en el caso anterior, para la elaboración de esta matriz se han filtrado los 8 distritos con un mayor volumen de viajes asociados. A partir de la figura podemos ver que destacan los viajes que tienen su origen y destino en el distrito de Carabanchel, probablemente debido a que dicho distrito es el más poblado de la capital. Asimismo, las conexiones entre el distrito Centro y el resto acumulan también una gran cantidad de viajes.

7.2.2. Mapas de frecuencia de viajes OD

Los mapas de frecuencia de viajes origen-destino son utilizados como medio auxiliar para representar sobre ellos un resumen del número de viajes que salen y llegan a una misma ubicación, a diferentes niveles de agrupación. Estos gráficos suplen la incapacidad de los mapas de calor anteriores para representar geográficamente la información de viajes.

Mapa de frecuencia de viajes OD agrupado por municipio. La Figura 7.33 muestra un mapa de frecuencia de viajes OD en los municipios más transitados de Madrid. Por medio de las flechas descritas en la leyenda se indica el número de viajes que salen y llegan a cada municipio representado en el mapa. Podemos observar en la figura como hay

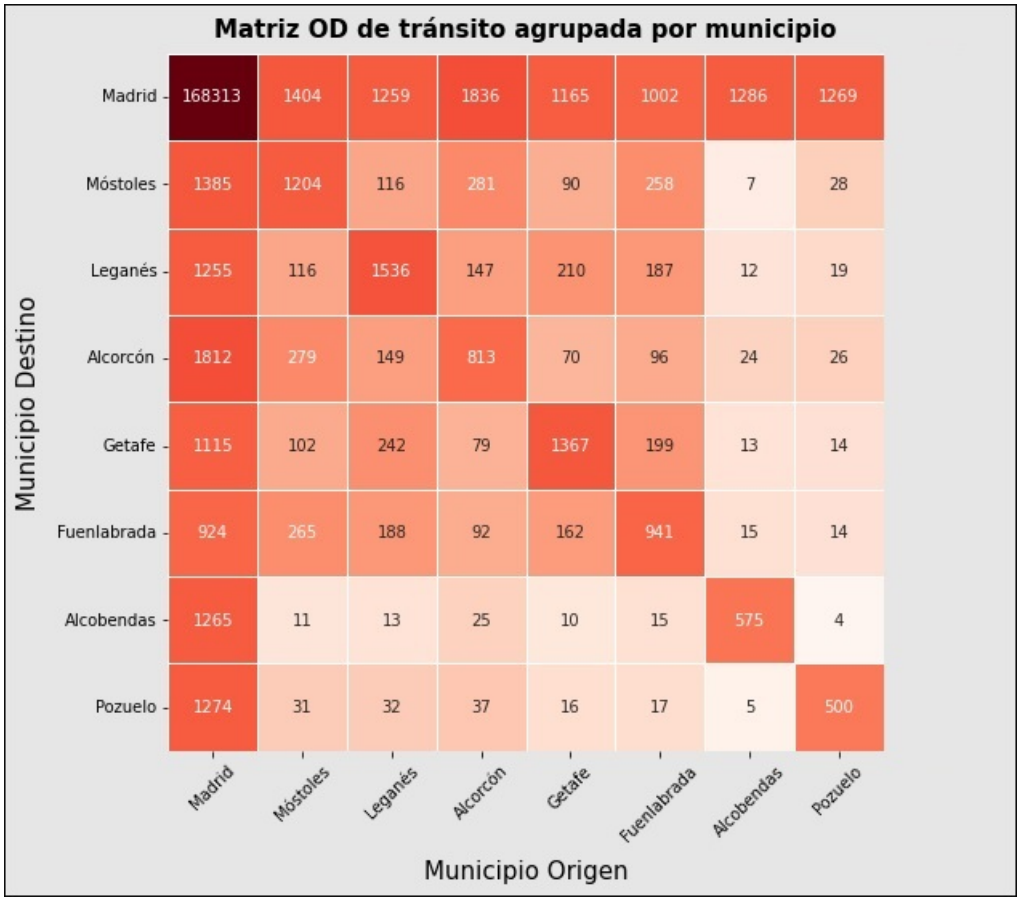

Figura 7.31: Matriz OD de tránsito agrupada por municipio.

poca variación entre el número de viajes que tienen como origen y destino cada uno de los municipios considerados.

Mapa de frecuencia de viajes OD agrupado por distrito de Madrid Capital. La Figura 7.34 muestra un mapa de frecuencia de viajes OD en los distritos más transitados de Madrid Capital. Como en el mapa anterior, las flechas indican el número de viajes que salen y llegan a cada uno de los distritos más frecuentados. Al contrario de lo deducido en el mapa de frecuencia referente a los municipios, en este caso sí puede observarse mayor variación entre el número de viajes que salen y llegan a cada distrito.

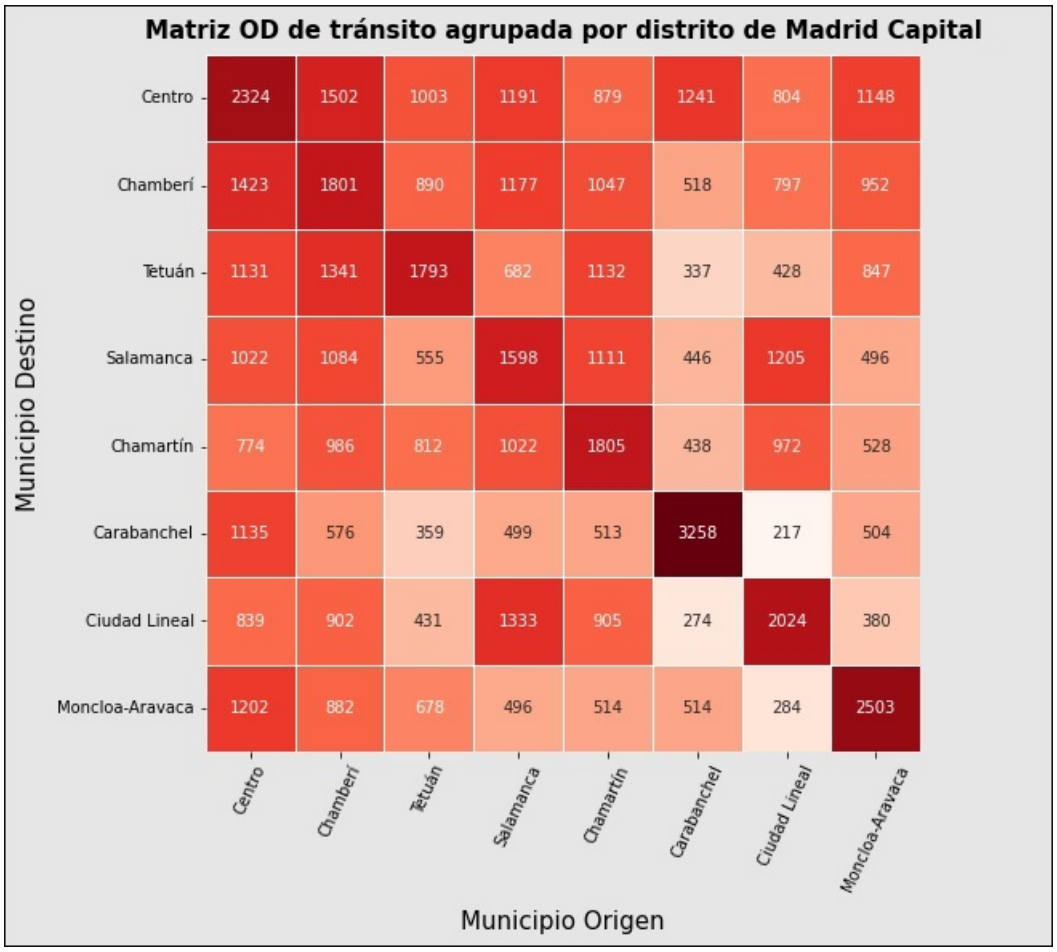

Figura 7.32: Matriz OD de tránsito agrupada por distrito de la ciudad de Madrid.

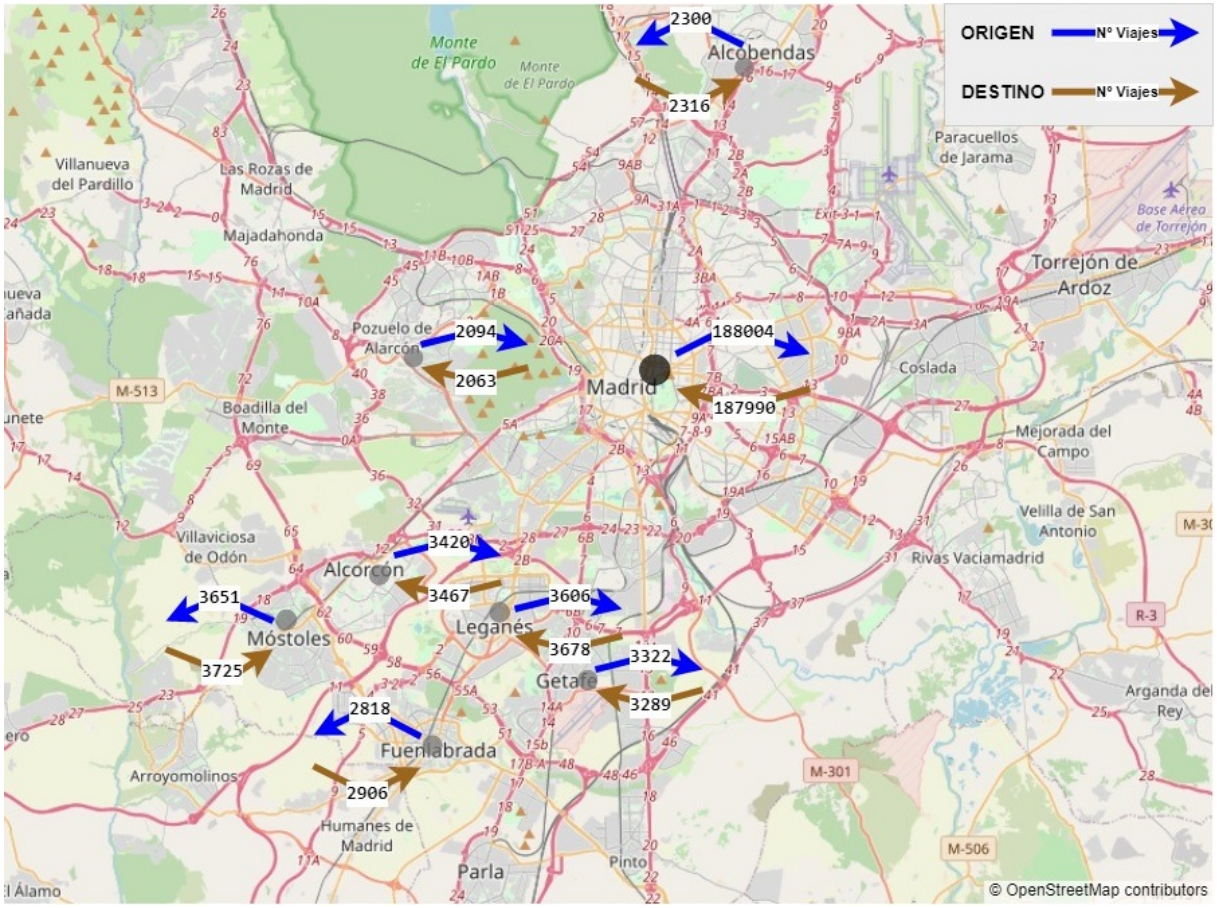

Figura 7.33: Volumen de viajes OD en los municipios más transitados de Madrid.

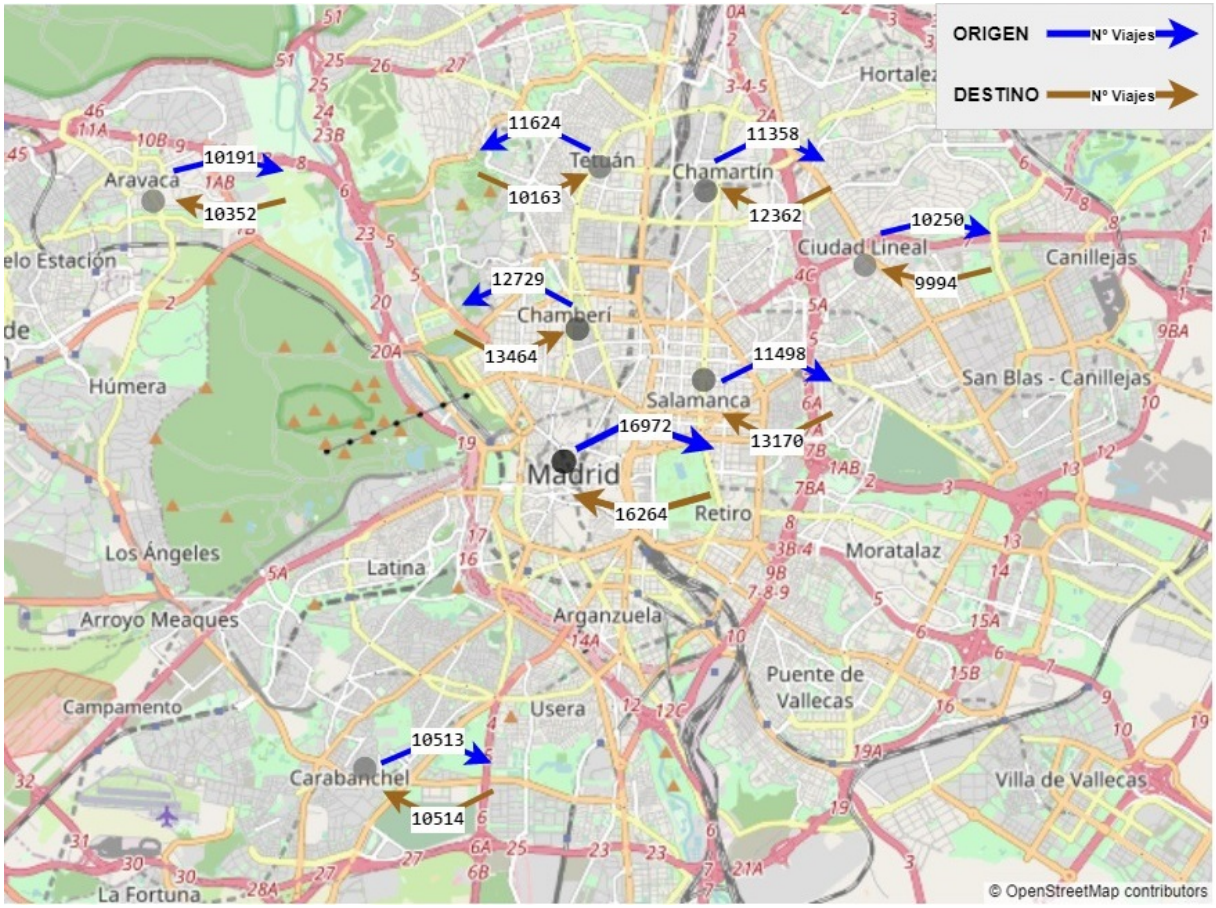

Figura 7.34: Volumen de viajes OD en los distritos más transitados de Madrid Capital.

Conclusiones y Trabajo Futuro

En este capítulo del proyecto se hace un análisis de las conclusiones más importantes que se infieren tanto a nivel de proyecto como de la aplicación del método *trip chaining* sobre la red de transporte público de Madrid. Además, se propondrán varias líneas de trabajo futuro para mejorar la solución actual del proyecto y considerar otras propuestas de investigación y desarrollo.

8.1. Conclusiones

A lo largo de este trabajo se han ido completando de forma iterativa e incremental las diferentes etapas del ciclo de vida que caracterizan un proyecto *Big Data* de acuerdo con una metodología de desarrollo adaptada del estándar CRISP-DM. Aunque el objetivo principal sea analizar la viabilidad de la aplicación del método *trip chaining* sobre la red de transporte público de Madrid, este proyecto ha ido más allá, condensando las etapas preliminares a la implementación de este método, constituyendo una solución completa. La realización de las tareas del proyecto se ha estructurado en torno a una serie de sprints de corta duración siguiendo la metodología de trabajo UVagile, derivada del marco de trabajo ágil Scrum.

En cuanto a los resultados obtenidos en el proyecto, se destaca la capacidad del método *trip chaining* para estimar las paradas de bajada de los pasajeros de la red de transporte de Madrid, lográndose una tasa de acierto del 81,97 % entre los segmentos de viaje cuya parada de bajada pudo ser validada por ser conocida. Asimismo, el método es capaz de inferir la bajada para un 89,60 % del total de segmentos. No obstante, se debe tener en cuenta que la validación realizada para evaluar el rendimiento del método ha sido parcial, siendo acometida solamente sobre una parte de los segmentos de viaje de los que teníamos información sobre su parada de bajada real. Además, se estableció un margen de error permitido de 1 parada de ruta de distancia a la hora de considerar si las estimaciones realizadas son correctas. Pese a todo, los resultados obtenidos en este proyecto inicial demuestran la viabilidad de la aplicación del método *trip chaining*.

Por otro lado, se han elaborado una serie de matrices OD de tránsito agrupadas sobre diferentes criterios que, aunque solo representan los viajes realizados para una muestra de 100.000 pasajeros de la red de transporte, permiten dar una visión resumida de la dinámica general de viajes que tiene lugar en la red de transporte de Madrid.

Por último, este proyecto constituye un estudio inicial que presenta posibilidades de mejora de cara a optimizar tanto las estadísticas de rendimiento como la presentación gráfica de los resultados obtenidos en el proyecto. Por ello, este proyecto puede ser considerado como un buen punto de partida para continuar la investigación y desarrollo en temas relacionados con el transporte urbano inteligente.

8.2. Trabajo Futuro

El trabajo futuro representa la oportunidad de continuar el desarrollo y la investigación del proyecto actual. Se abre la posibilidad de retomar el trabajo del proyecto desde su estado actual para emprender mejoras en su implementación y buscar nuevas contribuciones que aporten un valor añadido a la solución propuesta.

Tras analizar los resultados obtenidos en este proyecto, se derivan un conjunto de posibles líneas de trabajo futuro que pueden ser consideradas para mejorar los estadísticas de rendimiento actuales y proponer otras nuevas vías de investigación. Desde este punto de vista, se plantean 4 líneas de trabajo futuro que pueden considerarse interesantes para continuar trabajando en el proyecto.

Optimización de la ejecución del código del proyecto. Esta primera línea de trabajo pone el foco en acelerar el tiempo requerido para ejecutar la carga de trabajo (*workload*) completa implementada en el proyecto para procesar el gran volumen de datos de transacciones de viaje de los que disponemos. Para ello, se propone la migración del código de implementación a un entorno de ejecución distribuida en el que la carga de trabajo se pueda ejecutar en paralelo de forma distribuida, para así poder procesar un mayor volumen de datos y obtener unos resultados más representativos en menor tiempo.

Respecto a este proyecto, en el que el código está ya implementado en el lenguaje de programación Python, la solución más directa sería optar por utilizar Ray, un framework de procesamiento distribuido designado para trabajar sobre Python con el que agilizar la ejecución de cargas de trabajo computacionalmente intensivas. El objetivo perseguido con esta optimización es conseguir que se puedan ejecutar en un tiempo asumible una mayor cantidad de transacciones de viaje con las que obtener resultados finales más consistentes.

Elaboración de un *dashboard* para visualizar los resultados del proyecto. Esta línea de trabajo propuesta consistiría en hacer el diseño e implementación de un dashboard, también denominado cuadro de mandos, que englobe en una misma vista gráficos con los resultados más destacados del proyecto. La principal ventaja del *dashboard* frente a otras clásicas representaciones gráficas de datos radica en que ofrece una visión de los

datos dinámica e interactiva en la que los gráficos dispuestos pueden ir adaptándose a las consultas seleccionadas por parte del usuario. Así, la interfaz gráfica ofrecida se modificará en su conjunto para ajustar la información de sus gráficos, permitiendo al usuario conocer nuevos aspectos sobre los datos.

Investigación de otros métodos de estimación de las paradas de bajada. Esta línea de trabajo propone investigar la posibilidad de aplicar distintos métodos de estimación de las paradas de bajada en el ámbito de la red de transporte público de Madrid. En la literatura, se han planteado 3 aproximaciones principales de cara a la resolución de este problema de estimación.

Para este proyecto, la primera opción sería intentar mejorar la propuesta actual de aplicación del método *trip chaining*. Para ello, se podrían valorar distintas posibilidades, tales como la modificación de las suposiciones del método, la variación de los parámetros utilizados para la detección de transbordos o el uso de nuevas funciones auxiliares para tratar aspectos particulares de la red de transporte de Madrid. Los resultados de rendimiento obtenidos con cada aproximación de mejora probada deberán ser analizados de forma comparativa para seleccionar la opción más óptima. No obstante, esta adaptación del método no sería directa, pues requeriría llevar a cabo un análisis profundo de la topología y la dinámica de viajes que se producen en la red de transporte de Madrid, para obtener un mayor conocimiento que sustente la adopción de las decisiones finales. La otra opción sería iniciar la investigación del funcionamiento específico y emprender el desarrollo desde cero de otros métodos de estimación, tales como el probabilístico o el basado en *deep learning*, con el objetivo de conseguir una mejora de los resultados obtenidos con la solución actual del proyecto.

Obtención de fuentes de datos adicionales. Esta última línea de trabajo futuro se basa en la búsqueda de nuevos conjuntos de datos que puedan enriquecer los utilizados actualmente en proyecto de cara a poder tomar decisiones de implementación más adecuadas según las características concretas de cada caso particular en el que nos encontremos.

Entre los potenciales conjuntos de datos cuya obtención sería de interés para el proyecto destacan: los asociados a estadísticos de uso del transporte público, los relativos a hábitos de los diferentes perfiles de pasajeros en sus viajes por la red de transporte, los datos de densidad de tránsito por ubicación, los datos poblacionales por zonas, u otros datos generales de contexto de Madrid. Dentro de esta línea de trabajo también se incluiría la necesidad de disponer de datos de transacciones de viaje lo más consistentes posible, pues los disponibles para este proyecto presentan ciertas anomalías y ausencia de información importante que conlleva la eliminación de un gran número de ellos.

En general, el principal impedimento para obtener este tipo de conjuntos de datos radica en que comúnmente no están disponibles de forma abierta a cualquier persona, sino que suelen ser datos privados utilizados para usos corporativos internos. Además,

consideraciones relativas a temas legales, como la protección de datos, implican una serie de restricciones al acceso a estos datos que suponen una dificultad añadida para su uso.

Apéndices

Apéndice A

Diccionario de Datos

En este apéndice se muestran todas las tablas del diccionario de datos descrito en el Apartado 4.1.2 del presente documento.

Entidades

ESTACIÓN						
Definición	Representa cada una de las ubicaciones dentro de la red de transporte público donde los pasajeros pueden subir y bajar de los vehículos de transporte.					
Notas	Desde una misma estación se pueden seguir distintas rutas de transporte.					
Atributos	**Nombre**	**Definición**	**Tipo**	**UNIQUE**	**NULL**	**DEFAULT**
	Id	Identificador de la estación en la red de transporte.	STRING	SÍ	NO	-
	Nombre	Denominación de la estación.	STRING	NO	NO	-
	Municipio	Municipio al que pertenece la estación.	STRING	NO	NO	-
	Distrito	Distrito en el que se encuentra la estación.	STRING	NO	SÍ	-
	Corona	Zona tarifaria en la que se ubica la estación.	ENUM	NO	SÍ	-
	Latitud	Coordenada de la latitud donde se ubica la estación expresada en grados.	FLOAT	NO	NO	-
	Longitud	Coordenada de la longitud donde se ubica la estación expresada en grados.	FLOAT	NO	NO	-

Figura A.1: Tabla Entidad Estación - Diccionario de Datos.

RUTA						
Definición	Representa cada una de las rutas de viaje ofrecidas por los diferentes modos de la red de transporte público.					
Notas	Cada ruta tiene ya asociado un sentido de viaje en la línea determinada.					
Atributos	**Nombre**	**Definición**	**Tipo**	**UNIQUE**	**NULL**	**DEFAULT**
	Id	Identificador de la ruta de transporte.	STRING	SÍ	NO	-
	Modo	Identificador del modo de transporte encargado de la ruta de transporte.	ENUM	NO	NO	-
	Número_Línea	Línea a la que pertenece la ruta de transporte.	STRING	NO	NO	-
	Número_Sublínea	Sublínea dentro de la línea a la que pertenece la ruta de transporte.	STRING	NO	SÍ	-
	Sentido	Dirección seguida por la ruta de transporte.	ENUM	NO	NO	-

Figura A.2: Tabla Entidad Ruta - Diccionario de Datos.

PARADA_RUTA						
Definición	Representa cada una de las paradas por las que transcurre una ruta de transporte.					
Notas	Cada parada está asociada con una ruta y una estación de la red de transporte público.					
Atributos	**Nombre**	**Definición**	**Tipo**	**UNIQUE**	**NULL**	**DEFAULT**
	Id	Identificador de la parada, resultado de concatenar los identificadores de ruta y estación asociados.	STRING	SÍ	NO	-
	Tipo	Tipo de la parada dentro de la ruta.	ENUM	NO	NO	-
	Orden	Número de orden de la parada dentro de la ruta.	INT	NO	NO	-
	Dpaypoint	Código de los terminales de registro de transacciones asociados a la parada.	STRING	NO	SÍ	-

Figura A.3: Tabla Entidad Parada Ruta - Diccionario de Datos.

VIAJE						
Definición	Representa cada viaje completo realizado por un pasajero en el transporte público.					
Notas	Cada viaje tiene asociados un origen y un destino.					
Atributos	**Nombre**	**Definición**	**Tipo**	**UNIQUE**	**NULL**	**DEFAULT**
	Id	Identificador del viaje en el transporte público.	STRING	SÍ	NO	-
	Fecha_Hora_ Inicio	Fecha de inicio del viaje.	DATETIME	NO	NO	-
	Fecha_Hora_ Fin	Fecha de finalización del viaje.	DATETIME	NO	SÍ	-
	Tipo	Tipo del viaje: reconstruido o no reconstruido.	STRING	NO	NO	-
	Tarjeta	Tarjeta de transporte personal asociada a un pasajero.	STRING	NO	NO	-

Figura A.4: Tabla Entidad Viaje - Diccionario de Datos.

ETAPA						
Definición	Representa cada etapa de viaje realizada por un pasajero en el transporte público.					
Notas	Cada etapa tiene asociadas una subida y una bajada, independientemente de si constituyen un origen y destino de viaje.					
Atributos	**Nombre**	**Definición**	**Tipo**	**UNIQUE**	**NULL**	**DEFAULT**
	Id	Identificador de la etapa de viaje en el transporte público.	STRING	SÍ	NO	-
	Fecha_Hora_ Inicio	Fecha y hora de inicio de la etapa de viaje.	DATETIME	NO	NO	-
	Fecha_Hora_ Fin	Fecha y hora de finalización de la etapa de viaje.	DATETIME	NO	SÍ	-
	Tipo	Tipo de la bajada del segmento de viaje: transbordo, destino o inconexo.	STRING	NO	NO	-

Figura A.5: Tabla Entidad Etapa - Diccionario de Datos.

Relaciones

CONTENER			
Definición	Empareja cada parada de ruta con la estación en la que se encuentra.		
Notas			
Entidades	**Nombre**	**Participación**	**Cardinalidad**
	ESTACIÓN	1	N
	PARADA_RUTA	1	1

Figura A.6: Tabla Relación Contener - Diccionario de Datos.

TRANSCURRIR			
Definición	Relaciona cada ruta de transporte con las paradas por las que transcurre.		
Notas			
Entidades	**Nombre**	**Participación**	**Cardinalidad**
	RUTA	1	N
	PARADA_RUTA	1	1

Figura A.7: Tabla Relación Transcurrir - Diccionario de Datos.

INICIAR			
Definición	Relaciona cada etapa de viaje con su parada de ruta de subida.		
Notas			
Entidades	**Nombre**	**Participación**	**Cardinalidad**
	ETAPA	1	1
	PARADA_RUTA	0	N

Figura A.8: Tabla Relación Iniciar - Diccionario de Datos.

FINALIZAR			
Definición	Relaciona cada etapa de viaje con su parada de ruta de bajada.		
Notas	De una etapa puede no conocerse/asignarse su parada de bajada.		
Entidades	**Nombre**	**Participación**	**Cardinalidad**
	ETAPA	0	1
	PARADA_RUTA	0	N

Figura A.9: Tabla Relación Finalizar - Diccionario de Datos.

AGRUPAR			
Definición	Reúne conjuntos de etapas en torno a un mismo viaje.		
Notas	El origen y destino de un viaje se corresponden con la parada de subida de su primera etapa y la parada de bajada de su última etapa, respectivamente.		
Entidades	**Nombre**	**Participación**	**Cardinalidad**
	VIAJE	1	N
	ETAPA	1	1

Figura A.10: Tabla Relación Agrupar - Diccionario de Datos.

DISTANCIAR						
Definición	Relaciona, en términos de distancia geográfica, cada par de estaciones de la red de transporte público.					
Notas	La medida de distancia a utilizar será la distancia ortodrómica, en unidades de metro.					
Entidades	**Nombre**		**Participación**		**Cardinalidad**	
	ESTACIÓN		1		N	
	ESTACIÓN		1		N	
Atributos	**Nombre**	**Definición**	**Tipo**	**UNIQUE**	**NULL**	**DEFAULT**
	Distancia	Distancia, en metros, entre 2 estaciones de la red.	FLOAT	NO	NO	-

Figura A.11: Tabla Relación Distanciar - Diccionario de Datos.

TARDAR						
Definición	Relaciona, en términos de tiempo, cada par de paradas consecutivas dentro de una ruta de transporte.					
Notas	La unidad de medida del tiempo a utilizar será el minuto.					
Entidades	**Nombre**		**Participación**		**Cardinalidad**	
	PARADA_RUTA		1		N	
	PARADA_RUTA		1		N	
Atributos	**Nombre**	**Definición**	**Tipo**	**UNIQUE**	**NULL**	**DEFAULT**
	Tiempo	Tiempo, en minutos, entre 2 paradas consecutivas de una ruta.	FLOAT	NO	NO	-

Figura A.12: Tabla Relación Tardar - Diccionario de Datos.

Apéndice B

Perfiles de Datos

En este apéndice se muestran todas las tablas de perfiles de datos referidas en el Apartado 4.2.1 del presente documento.

Datos del Portal de Datos Abiertos del CRTM

M4 ESTACIONES	
Definición	Contiene información descriptiva acerca de las estaciones de **Metro** de la red de transporte público de Madrid.
Fuente	Portal de Datos Abiertos del CRTM.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero CSV (campos separados por ",").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	30 de noviembre de 2018.
Anomalías	- Presencia de campos completamente vacíos. - Expresión de las coordenadas geográficas en un sistema de coordenadas (UTM ED50 Zona 30 Norte) distinto al de GPS.
Decisiones	- Descartar campos vacíos. - Transformación de la latitud y longitud al sistema de coordenadas utilizado en GPS (WGS84).

Figura B.1: Tabla Primera de M4 Estaciones - Perfiles de Datos.

M4 ESTACIONES		
Campo	**Descripción**	**Tipo de Datos**
IDESTACION	Identificador de la estación dentro de la red de transporte público de Madrid.	STRING
DENOMINACION	Nombre asignado a la estimación.	STRING
CODIGOMUNICIPIO	Código del municipio al que pertenece la estación.	STRING
DISTRITO	Código del distrito en el que se encuentra la estación.	STRING
CORONATARIFARIA	Código asociado a la zona tarifaria en la que se ubica la estación.	STRING
X	Coordenada X asociada a la estación según el sistema de coordenadas UTM ED50 Zona 30 Norte.	INT(6)
Y	Coordenada Y asociada a la estación según el sistema de coordenadas UTM ED50 Zona 30 Norte.	INT(7)
Otros atributos: OBJECTID, FECHAACTUAL, MODO, CODIGOESTACION, OBSERVACIONES, SITUACION, CODIGOCTMESTACIONREDMETRO, CODIGOEMPRESA, DENOMINACIONABREVIADA, MODOINTERCAMBIADOR, CODIGOINTERCAMBIADOR, TIPO, CODIGOPROVINCIA, CODIGOENTIDAD, CODIGONUCLEO, CODIGOVIA, TIPOVIA, PARTICULA, NOMBREVIA, TIPONUMERO, NUMEROPORTAL, CALIFICADORPORTAL, CARRETERA, CODIGOPOSTAL, SECCIONCENSAL, BARRIO, TESELA, SECTORURBANO, SECTOR, CORREDOR, CORONA123, ZONATRANSPORTE, ENCUESTADOMICILIARIA, ENCUESTAAFOROS, HOJA25000, ACONDICIONAMIENTOVIAJEROS, ACONDICIONAMIENTOVEHICULOS, FECHAALTA, FECHAINICIO, FECHAFIN, GRADOACCESIBILIDAD, SITUACIONCALLE, DENOMINACION_SAE, INTERURBANOS_CODIGOEMT_CRTM, INTERURBANOS_CODIGOEMT_EMPRESA.		

Figura B.2: Tabla Segunda de M4 Estaciones - Perfiles de Datos.

M4 ESTACIONES							
Campo	**Cardinalidad**	**Densidad**	**Rango**				**Formato**
			MIN	**MAX**	**AVG**	**STD**	
IDESTACION	289	100%	4_1	4_345	-	-	4_[0-9]{1,3}
DENOMINACION	241	100%	-	-	-	-	-
CODIGOMUNICIPIO	12	100%	006	134	-	-	[0-9]{3}
DISTRITO	21	100%	01	21	-	-	[0-9]{2}
CORONATARIFARIA	4	100%	-	-	-	-	A|B1|B2|B3
X	268	100%	425790	462079	441668,42	5340,28	[0-9]{6}
Y	273	100%	4459646	4490276	4474492,86	5599,78	[0-9]{7}
Número Total de Registros							**289**

Figura B.3: Tabla Tercera de M4 Estaciones - Perfiles de Datos.

M4 TRAMOS	
Definición	Almacena información descriptiva sobre los diferentes tramos, caracterizados por sus paradas terminantes, que componen cada una de las rutas de transporte ofrecidas en el **Metro**.
Fuente	Portal de Datos Abiertos del CRTM.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero CSV (campos separados por ",").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	30 de noviembre de 2018.
Anomalías	- Presencia de campos completamente vacíos.
Decisiones	- Descartar campos vacíos.

Figura B.4: Tabla Primera de M4 Tramos - Perfiles de Datos.

M10 TRAMOS		
Campo	**Descripción**	**Tipo de Datos**
CODIGOITINERARIO	Identificador de la ruta de transporte dentro de la red.	STRING
MODO	Modo de transporte en el que se transcurre la ruta.	STRING
NUMEROLINEAUSUARIO	Denominación de la línea a la que pertenece la ruta.	STRING
SENTIDO	Dirección de la ruta dentro de la línea.	STRING
NUMEROORDEN	Número de orden de la parada en la ruta.	INT
TIPOPARADA	Carácter correspondiente al tipo de la parada en la ruta.	STRING
LONGITUDTRAMOANTERIOR	Longitud, en metros, del tramo de ruta que termina en la parada.	FLOAT
VELOCIDADTRAMOANTERIOR	Velocidad media, en kilómetros por hora, a la que se recorre el tramo de ruta que termina en la parada.	FLOAT
IDFESTACION	Identificador de la estación dentro de la red asociada a la parada.	STRING
Otros atributos: OBJECTID, IDTRAMO, FECHAACTUAL, CODIGOGESTIONLINEA, TIPOITINERARIO, CODIGOESTACION, CODIGOPOSTE, CODIGOANDEN, IDENTIFICADORTIPOPARADA, DENOMINACION, CODIGOPROVINCIA, CODIGOMUNICIPIO, MUNICIPIO, CORONATARIFARIA, DIRECCION, FECHAALTA, FECHAINICIO, FECHAFIN, MODOLINEA, MODOINTERCAMBIADOR, CODIGOINTERCAMBIADOR, CODPROV_LINEA, CODMUN_LINEA, IDFTRAMO, CODIGOOBSERVACION, CODIGOSUBLINEA, DENOMINACION_SAE, IDFLINEA, IDFITINERARIO, IDFPOSTE, IDFANDEN, SHAPE_Length.		

Figura B.5: Tabla Segunda de M4 Tramos - Perfiles de Datos.

M4 TRAMOS							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
CODIGOITINERARIO	32	100%	283064	354330	-	-	[0-9]{6}
MODO	1	100%	4	4	-	-	4
NUMEROLINEAUSUARIO	18	100%	-	-	-	-	([0-9]{1,2}-?([12]|[ABab])?|R)
SENTIDO	2	100%	1	2	-	-	1|2
NUMEROORDEN	32	100%	2	33			[2-33]
TIPOPARADA	3	100%	-	-	-	-	C|I|T
LONGITUDTRAMOANTERIOR	556	100%	296,63	5750,27	1001,71	682,10	[0-9]{3,4},[0-9]+
VELOCIDADTRAMOANTERIOR	18	100%	19,00	54,34	28,41	7,00	[0-9]{2},[0-9]+
IDFESTACION	289	100%	4_1	4_345	-	-	4_[0-9]{1,2}
						Número Total de Registros	556

Figura B.6: Tabla Tercera de M4 Tramos - Perfiles de Datos.

M10 ESTACIONES	
Definición	Contiene información descriptiva acerca de las estaciones de **Metro Ligero** de la red de transporte público de Madrid.
Fuente	Portal de Datos Abiertos del CRTM.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero CSV (campos separados por ",").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	22 de junio de 2016.
Anomalías	- Presencia de campos completamente vacíos. - Expresión de las coordenadas geográficas en un sistema de coordenadas (UTM ED50 Zona 30 Norte) distinto al de GPS.
Decisiones	- Descartar campos vacíos. - Transformación de la latitud y longitud al sistema de coordenadas utilizado en GPS (WGS84).

Figura B.7: Tabla Primera de M10 Estaciones - Perfiles de Datos.

M10 ESTACIONES		
Campo	**Descripción**	**Tipo de Datos**
IDESTACION	Identificador de la estación dentro de la red de transporte público de Madrid.	STRING
DENOMINACION	Nombre asignado a la estimación.	STRING
CODIGOMUNICIPIO	Código del municipio al que pertenece la estación.	STRING
DISTRITO	Código del distrito en el que se encuentra la estación.	STRING
CORONATARIFARIA	Código asociado a la zona tarifaria en la que se ubica la estación.	STRING
X	Coordenada X asociada a la estación según el sistema de coordenadas UTM ED50 Zona 30 Norte.	INT(6)
Y	Coordenada Y asociada a la estación según el sistema de coordenadas UTM ED50 Zona 30 Norte.	INT(7)
Otros atributos: OBJECTID, FECHAACTUAL, MODO, CODIGOESTACION, OBSERVACIONES, SITUACION, CODIGOCTMESTACIONREDMETRO, CODIGOEMPRESA, DENOMINACIONABREVIADA, MODOINTERCAMBIADOR, CODIGOINTERCAMBIADOR, TIPO, CODIGOPROVINCIA, CODIGOENTIDAD, CODIGONUCLEO, CODIGOVIA, TIPOVIA, PARTICULA, NOMBREVIA, TIPONUMERO, NUMEROPORTAL, CALIFICADORPORTAL, CARRETERA, CODIGOPOSTAL, SECCIONCENSAL, BARRIO, TESELA, SECTORURBANO, SECTOR, CORREDOR, CORONA123, ZONATRANSPORTE, ENCUESTADOMICILIARIA, ENCUESTAAFOROS, HOJA25000, ACONDICIONAMIENTOVIAJEROS, ACONDICIONAMIENTOVEHICULOS, FECHAALTA, FECHAINICIO, FECHAFIN, GRADOACCESIBILIDAD, SITUACIONCALLE, DENOMINACION_SAE, INTERURBANOS_CODIGOEMT_CRTM, INTERURBANOS_CODIGOEMT_EMPRESA.		

Figura B.8: Tabla Segunda de M10 Estaciones - Perfiles de Datos.

M10 ESTACIONES							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
IDESTACION	57	100%	10_1	10_68	-	-	10_[0-9]{1,2}
DENOMINACION	56	100%	-	-	-	-	-
CODIGOMUNICIPIO	5	100%	007	115	-	-	[0-9]{3}
DISTRITO	7	100%	01	16	-	-	[0-9]{2}
CORONATARIFARIA	3	100%	-	-	-	-	A|B1|B2
X	57	100%	423423	444908	434322,77	5475,15	[0-9]{6}
Y	56	100%	4452872	4484592	4468678,56	10951,62	[0-9]{7}
						Número Total de Registros	57

Figura B.9: Tabla Tercera de M10 Estaciones - Perfiles de Datos.

M10 TRAMOS	
Definición	Almacena información descriptiva sobre los diferentes tramos, caracterizados por sus paradas terminantes, que componen cada una de las rutas de transporte ofrecidas en el **Metro Ligero**.
Fuente	Portal de Datos Abiertos del CRTM.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero CSV (campos separados por ",").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	22 de junio de 2016.
Anomalías	- Presencia de campos completamente vacíos.
Decisiones	- Descartar campos vacíos.

Figura B.10: Tabla Primera de M10 Tramos - Perfiles de Datos.

M10 TRAMOS		
Campo	**Descripción**	**Tipo de Datos**
CODIGOITINERARIO	Identificador de la ruta de transporte dentro de la red.	STRING
MODO	Modo de transporte en el que se transcurre la ruta.	STRING
NUMEROLINEAUSUARIO	Denominación de la línea a la que pertenece la ruta.	STRING
SENTIDO	Dirección de la ruta dentro de la línea.	STRING
NUMEROORDEN	Número de orden de la parada en la ruta.	INT
TIPOPARADA	Carácter correspondiente al tipo de la parada en la ruta.	STRING
LONGITUDTRAMOANTERIOR	Longitud, en metros, del tramo de ruta que termina en la parada.	FLOAT
VELOCIDADTRAMOANTERIOR	Velocidad media, en kilómetros por hora, a la que se recorre el tramo de ruta que termina en la parada.	FLOAT
IDFESTACION	Identificador de la estación dentro de la red asociada a la parada.	STRING
Otros atributos: OBJECTID, IDTRAMO, FECHAACTUAL, CODIGOGESTIONLINEA, TIPOITINERARIO, CODIGOESTACION, CODIGOPOSTE, CODIGOANDEN,		

Figura B.11: Tabla Segunda de M10 Tramos - Perfiles de Datos.

M10 TRAMOS							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
CODIGOITINERARIO	32	100%	335989	339043	-	-	[0-9]{6}
MODO	1	100%	10	10	-	-	10
NUMEROLINEAUSUARIO	18	100%	-	-	-	-	[0-9](-[0-9])?
SENTIDO	2	100%	1	2	-	-	1|2
NUMEROORDEN	32	100%	2	16			[2-16]
TIPOPARADA	3	100%	-	-	-	-	C|I|T
LONGITUDTRAMOANTERIOR	556	100%	337,60	1846,58	706,34	344,04	[0-9]{3,4},[0-9]+
VELOCIDADTRAMOANTERIOR	18	100%	21,00	25,00	22,92	3,27	[0-9]{2},[0-9]+
IDFESTACION	57	100%	10_1	10_68	-	-	10_[0-9]{1,2}
						Número Total de Registros	100

Figura B.12: Tabla Tercera de M10 Tramos - Perfiles de Datos.

M5 ESTACIONES	
Definición	Contiene información descriptiva acerca de las estaciones de **Cercanías** de la red de transporte público de Madrid.
Fuente	Portal de Datos Abiertos del CRTM.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero CSV (campos separados por ",").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	25 de septiembre de 2019.
Anomalías	- Presencia de campos completamente vacíos. - Expresión de las coordenadas geográficas en un sistema de coordenadas (UTM ED50 Zona 30 Norte) distinto al de GPS. - Presencia de valores nulos en los campos DISTRITO y CORONATARIFARIA.
Decisiones	- Descartar campos vacíos. - Transformación de la latitud y longitud al sistema de coordenadas utilizado en GPS (WGS84). - Mantener los valores nulos en los campos DISTRITO y CORONATARIFARIA.

Figura B.13: Tabla Primera de M5 Estaciones - Perfiles de Datos.

M5 ESTACIONES		
Campo	**Descripción**	**Tipo de Datos**
IDESTACION	Identificador de la estación dentro de la red de transporte público de Madrid.	STRING
DENOMINACION	Nombre asignado a la estimación.	STRING
CODIGOMUNICIPIO	Código del municipio al que pertenece la estación.	STRING
DISTRITO	Código del distrito en el que se encuentra la estación.	STRING
CORONATARIFARIA	Código asociado a la zona tarifaria en la que se ubica la estación.	STRING
X	Coordenada X asociada a la estación según el sistema de coordenadas UTM ED50 Zona 30 Norte.	INT(6)
Y	Coordenada Y asociada a la estación según el sistema de coordenadas UTM ED50 Zona 30 Norte.	INT(7)
Otros atributos: OBJECTID, FECHAACTUAL, MODO, CODIGOESTACION, OBSERVACIONES, SITUACION, CODIGOCTMESTACIONREDMETRO, CODIGOEMPRESA, DENOMINACIONABREVIADA, MODOINTERCAMBIADOR, CODIGOINTERCAMBIADOR, TIPO, CODIGOPROVINCIA, CODIGOENTIDAD, CODIGONUCLEO, CODIGOVIA, TIPOVIA, PARTICULA, NOMBREVIA, TIPONUMERO, NUMEROPORTAL, CALIFICADORPORTAL, CARRETERA, CODIGOPOSTAL, SECCIONCENSAL, BARRIO, TESELA, SECTORURBANO, SECTOR, CORREDOR, CORONA123, ZONATRANSPORTE, ENCUESTADOMICILIARIA, ENCUESTAAFOROS, HOJA25000, ACONDICIONAMIENTOVIAJEROS, ACONDICIONAMIENTOVEHICULOS, FECHAALTA, FECHAINICIO, FECHAFIN, GRADOACCESIBILIDAD, SITUACIONCALLE, DENOMINACION_SAE, INTERURBANOS_CODIGOEMT_CRTM, INTERURBANOS_CODIGOEMT_EMPRESA, LINEAS.		

Figura B.14: Tabla Segunda de M5 Estaciones - Perfiles de Datos.

M5 ESTACIONES							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
IDESTACION	111	100%	5_1	5_145	-	-	5_[0-9]{1,3}
DENOMINACION	111	100%	-	-	-	-	-
CODIGOMUNICIPIO	47	100%	005	904	-	-	[0-9]{3}
DISTRITO	7	96,4%	01	21	-	-	[0-9]{2}
CORONATARIFARIA	8	86,5%	-	-	-	-	A|B1|B2|B3|C1|C2|E1|SZ|
X	111	100%	340254	578832	433248,19	32744,47	[0-9]{6}
Y	111	100%	4413913	4611695	4484763,91	27438,81	[0-9]{7}
						Número Total de Registros	111

Figura B.15: Tabla Tercera de M5 Estaciones - Perfiles de Datos.

M5 TRAMOS	
Definición	Almacena información descriptiva sobre los diferentes tramos, caracterizados por sus paradas terminantes, que componen cada una de las rutas de transporte ofrecidas en el **Cercanías**.
Fuente	Portal de Datos Abiertos del CRTM.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero CSV (campos separados por ",").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	25 de septiembre de 2019.
Anomalías	- Presencia de campos completamente vacíos.
Decisiones	- Descartar campos vacíos.

Figura B.16: Tabla Primera de M5 Tramos - Perfiles de Datos.

M5 TRAMOS		
Campo	**Descripción**	**Tipo de Datos**
CODIGOITINERARIO	Identificador de la ruta de transporte dentro de la red.	STRING
MODO	Modo de transporte en el que se transcurre la ruta.	STRING
NUMEROLINEAUSUARIO	Denominación de la línea a la que pertenece la ruta.	STRING
SENTIDO	Dirección de la ruta dentro de la línea.	STRING
NUMEROORDEN	Número de orden de la parada en la ruta.	INT
TIPOPARADA	Carácter correspondiente al tipo de la parada en la ruta.	STRING
LONGITUDTRAMOANTERIOR	Longitud, en metros, del tramo de ruta que termina en la parada.	FLOAT
VELOCIDADTRAMOANTERIOR	Velocidad media, en kilómetros por hora, a la que se recorre el tramo de ruta que termina en la parada.	FLOAT
CODIGOESTACION	Identificador de la estación dentro de la red de **Cercanías**.	STRING
Otros atributos: OBJECTID, IDTRAMO, FECHAACTUAL, CODIGOGESTIONLINEA, TIPOITINERARIO, CODIGOPOSTE, CODIGOANDEN, IDENTIFICADORTIPOPARADA, DENOMINACION, CODIGOPROVINCIA, CODIGOMUNICIPIO, MUNICIPIO, CORONATARIFARIA, DIRECCION, FECHAALTA, FECHAINICIO, FECHAFIN, MODOLINEA, MODOINTERCAMBIADOR, CODIGOINTERCAMBIADOR, CODPROV_LINEA, CODMUN_LINEA, IDFTRAMO, CODIGOOBSERVACION, CODIGOSUBLINEA, DENOMINACION_SAE, IDFLINEA, SHAPE_Length.		

Figura B.17: Tabla Segunda de M5 Tramos - Perfiles de Datos.

M5 TRAMOS							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
CODIGOITINERARIO	22	100%	330338	364229	-	-	[0-9]{6}
MODO	1	100%	5	5	-	-	5
NUMEROLINEAUSUARIO	11	100%	-	-	-	-	C-[0-9]([ab]|[0-9])?
SENTIDO	2	100%	1	2	-	-	1|2
NUMEROORDEN	31	100%	2	32			[2-32]
TIPOPARADA	3	100%	-	-	-	-	C|I|T
LONGITUDTRAMOANTERIOR	244	100%	590,15	15458,39	3690,04	2715,08	[0-9]{3,5},[0-9]+
VELOCIDADTRAMOANTERIOR	1	100%	30,00	30,00	30,00	30,00	[0-9]{2},[0-9]+
CODIGOESTACION	95	100%	1	145	-	-	[0-9]{1,3}
						Número Total de Registros	384

Figura B.18: Tabla Tercera de M5 Tramos - Perfiles de Datos.

M6 ESTACIONES	
Definición	Contiene información descriptiva acerca de las paradas de **Autobuses EMT** de la red de transporte público de Madrid.
Fuente	Portal de Datos Abiertos del CRTM.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero CSV (campos separados por ",").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	12 de septiembre de 2019.
Anomalías	- Presencia de campos completamente vacíos. - Expresión de las coordenadas geográficas en un sistema de coordenadas (UTM ED50 Zona 30 Norte) distinto al de GPS. - Presencia de valores nulos en los campos X e Y.
Decisiones	- Descartar campos vacíos. - Transformación de la latitud y longitud al sistema de coordenadas utilizado en GPS (WGS84). - Imputación manual de la longitud y latitud en los registros con valores nulos en X e Y.

Figura B.19: Tabla Primera de M6 Estaciones - Perfiles de Datos.

M6 ESTACIONES		
Campo	**Descripción**	**Tipo de Datos**
IDESTACION	Identificador de la estación dentro de la red de transporte público de Madrid.	STRING
DENOMINACION	Nombre asignado a la estimación.	STRING
CODIGOMUNICIPIO	Código del municipio al que pertenece la estación.	STRING
DISTRITO	Código del distrito en el que se encuentra la estación.	STRING
CORONATARIFARIA	Código asociado a la zona tarifaria en la que se ubica la estación.	STRING
X	Coordenada X asociada a la estación según el sistema de coordenadas UTM ED50 Zona 30 Norte.	INT(6)
Y	Coordenada Y asociada a la estación según el sistema de coordenadas UTM ED50 Zona 30 Norte.	INT(7)
Otros atributos: OBJECTID, FECHAACTUAL, MODO, CODIGOESTACION, OBSERVACIONES, SITUACION, CODIGOCTMESTACIONREDMETRO, CODIGOEMPRESA, DENOMINACIONABREVIADA, MODOINTERCAMBIADOR, CODIGOINTERCAMBIADOR, TIPO, CODIGOPROVINCIA, CODIGOENTIDAD, CODIGONUCLEO, CODIGOVIA, TIPOVIA, PARTICULA, NOMBREVIA, TIPONUMERO, NUMEROPORTAL, CALIFICADORPORTAL, CARRETERA, CODIGOPOSTAL, SECCIONCENSAL, BARRIO, TESELA, SECTORURBANO, SECTOR, CORREDOR, CORONA123, ZONATRANSPORTE, ENCUESTADOMICILIARIA, ENCUESTAAFOROS, HOJA25000, ACONDICIONAMIENTOVIAJEROS, ACONDICIONAMIENTOVEHICULOS, FECHAALTA, FECHAINICIO, FECHAFIN, GRADOACCESIBILIDAD, SITUACIONCALLE, DENOMINACION_SAE, INTERURBANOS_CODIGOEMT_CRTM, INTERURBANOS_CODIGOEMT_EMPRESA, LINEAS.		

Figura B.20: Tabla Segunda de M6 Estaciones - Perfiles de Datos.

M6 ESTACIONES							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
IDESTACION	4722	100%	6_10	6_6579	-	-	6_[0-9]{2,4}
DENOMINACION	3496	100%	-	-	-	-	-
CODIGOMUNICIPIO	2	100%	079	115	-	-	[0-9]{3}
DISTRITO	21	100%	01	21	-	-	[0-9]{2}
CORONATARIFARIA	2	100%	-	-	-	-	A|B1
X	111	99,8%	429206	454088	442208,16	4017,36	[0-9]{6}
Y	111	99,8%	4465068	4485580	4475336,47	4422,81	[0-9]{7}
						Número Total de Registros	4722

Figura B.21: Tabla Tercera de M6 Estaciones - Perfiles de Datos.

M6 TRAMOS	
Definición	Almacena información descriptiva sobre los diferentes tramos, caracterizados por sus paradas terminantes, que componen cada una de las rutas de transporte ofrecidas en **Autobuses EMT**.
Fuente	Portal de Datos Abiertos del CRTM.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero CSV (campos separados por ",").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	12 de marzo de 2021.
Anomalías	- Presencia de campos completamente vacíos. - Valores incorrectos en el campo TIPOPARADA. - Valores nulos e incorrectos en el campo VELOCIDADTRAMOANTERIOR.
Decisiones	- Descartar campos vacíos. - Imputación manual de los valores incorrectos en TIPOPARADA. - Determinación de los valores nulos e incorrectos de VELOCIDADTRAMOANTERIOR a partir de la media calculada con el resto de los registros.

Figura B.22: Tabla Primera de M6 Tramos - Perfiles de Datos.

M6 TRAMOS		
Campo	**Descripción**	**Tipo de Datos**
CODIGOITINERARIO	Identificador de la ruta de transporte dentro de la red.	STRING
MODO	Modo de transporte en el que se transcurre la ruta.	STRING
NUMEROLINEAUSUARIO	Denominación de la línea a la que pertenece la ruta.	STRING
SENTIDO	Dirección de la ruta dentro de la línea.	STRING
NUMEROORDEN	Número de orden de la parada en la ruta.	INT
TIPOPARADA	Carácter correspondiente al tipo de la parada en la ruta.	STRING
LONGITUDTRAMOANTERIOR	Longitud, en metros, del tramo de ruta que termina en la parada.	FLOAT
VELOCIDADTRAMOANTERIOR	Velocidad media, en kilómetros por hora, a la que se recorre el tramo de ruta que termina en la parada.	FLOAT
CODIGOESTACION	Identificador de la estación dentro de la red de **Autobuses EMT**.	STRING
Otros atributos: OBJECTID, IDTRAMO, FECHAACTUAL, CODIGOGESTIONLINEA, TIPOITINERARIO, CODIGOPOSTE, CODIGOANDEN, IDENTIFICADORTIPOPARADA, DENOMINACION, CODIGOPROVINCIA, CODIGOMUNICIPIO, MUNICIPIO, CORONATARIFARIA, DIRECCION, FECHAALTA, FECHAINICIO, FECHAFIN, MODOLINEA, MODOINTERCAMBIADOR, CODIGOINTERCAMBIADOR, CODPROV_LINEA, CODMUN_LINEA, IDFTRAMO, CODIGOOBSERVACION, CODIGOSUBLINEA, DENOMINACION_SAE, IDFLINEA, SHAPE_Length.		

Figura B.23: Tabla Segunda de M6 Tramos - Perfiles de Datos.

M6 TRAMOS							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
CODIGOITINERARIO	422	100%	283642	363948	-	-	[0-9]{6}
MODO	1	100%	6	6	-	-	6
NUMEROLINEAUSUARIO	211	100%	-	-	-	-	SE7[0-9]{2}|[A-Z][0-9]{0,2}|[0-9]{1,3}(SF)?
SENTIDO	2	100%	1	2	-	-	1|2
NUMEROORDEN	44	100%	2	45			[2-45]
TIPOPARADA	7	100%	-	-	-	-	C|I|T|IN|N|D|TN
LONGITUDTRAMOANTERIOR	6744	100%	50,00	12173,95	364,08	451,19	[0-9]{2,5},[0-9]+
VELOCIDADTRAMOANTERIOR	1331	98%	-1	40,75	12,92	10,34	[0-9]{2},[0-9]+
CODIGOESTACION	4663	100%	2	6581	-	-	[0-9]{1,4}
						Número Total de Registros	10439

Figura B.24: Tabla Tercera de M6 Tramos - Perfiles de Datos.

M8 ESTACIONES	
Definición	Contiene información descriptiva acerca de las paradas de **Autobuses Interurbanos** de la red de transporte público de Madrid.
Fuente	Portal de Datos Abiertos del CRTM.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero CSV (campos separados por ",").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	12 de septiembre de 2019.
Anomalías	- Presencia de campos completamente vacíos. - Expresión de las coordenadas geográficas en un sistema de coordenadas (UTM ED50 Zona 30 Norte) distinto al de GPS. - Presencia de un valor incorrecto, igual a "+", en el campo DENOMINACION. - Presencia de valores nulos en los campos DISTRITO y CORONATARIFARIA.
Decisiones	- Descartar campos vacíos. - Transformación de la latitud y longitud al sistema de coordenadas utilizado en GPS (WGS84). - Imputación del valor incorrecto en el campo DENOMINACION por el existente en el campo DENOMINACION_SAE ("AZUQUECA-COLEGIO"). - Mantener los valores nulos en los campos DISTRITO y CORONATARIFARIA.

Figura B.25: Tabla Primera de M8 Estaciones - Perfiles de Datos.

M8 ESTACIONES		
Campo	**Descripción**	**Tipo de Datos**
IDESTACION	Identificador de la estación dentro de la red de transporte público de Madrid.	STRING
DENOMINACION	Nombre asignado a la estimación.	STRING
CODIGOMUNICIPIO	Código del municipio al que pertenece la estación.	STRING
DISTRITO	Código del distrito en el que se encuentra la estación.	STRING
CORONATARIFARIA	Código asociado a la zona tarifaria en la que se ubica la estación.	STRING
X	Coordenada X asociada a la estación según el sistema de coordenadas UTM ED50 Zona 30 Norte.	INT(6)
Y	Coordenada Y asociada a la estación según el sistema de coordenadas UTM ED50 Zona 30 Norte.	INT(7)
Otros atributos: OBJECTID, FECHAACTUAL, MODO, CODIGOESTACION, OBSERVACIONES, SITUACION, CODIGOCTMESTACIONREDMETRO, CODIGOEMPRESA, DENOMINACIONABREVIADA, MODOINTERCAMBIADOR, CODIGOINTERCAMBIADOR, TIPO, CODIGOPROVINCIA, CODIGOENTIDAD, CODIGONUCLEO, CODIGOVIA, TIPOVIA, PARTICULA, NOMBREVIA, TIPONUMERO, NUMEROPORTAL, CALIFICADORPORTAL, CARRETERA, CODIGOPOSTAL, SECCIONCENSAL, BARRIO, TESELA, SECTORURBANO, SECTOR, CORREDOR, CORONA123, ZONATRANSPORTE, ENCUESTADOMICILIARIA, ENCUESTAAFOROS, HOJA25000, ACONDICIONAMIENTOVIAJEROS, ACONDICIONAMIENTOVEHICULOS, FECHAALTA, FECHAINICIO, FECHAFIN, GRADOACCESIBILIDAD, SITUACIONCALLE, DENOMINACION_SAE, INTERURBANOS_CODIGOEMT_CRTM, INTERURBANOS_CODIGOEMT_EMPRESA, LINEAS.		

Figura B.26: Tabla Segunda de M8 Estaciones - Perfiles de Datos.

M8 ESTACIONES							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
IDESTACION	8493	100%	8_06002	8_99987	-	-	8_[0-9]{5}
DENOMINACION	6244	100%	-	-	-	-	-
CODIGOMUNICIPIO	209	100%	001	903	-	-	[0-9]{3}
DISTRITO	22	96,1%	0	21	-	-	[0-9]{1,2}
CORONATARIFARIA	10	99,9%	-	-	-	-	A|B1|B2|B3|C1|C2|E1|E2|Ex|SZ
X	7883	100%	365798	573855	438497,92	19889,14	[0-9]{6}
Y	7906	100%	4331220	4553734	4475431.40	20609.75	[0-9]{7}
						Número Total de Registros	8493

Figura B.27: Tabla Tercera de M8 Estaciones - Perfiles de Datos.

M8 TRAMOS	
Definición	Almacena información descriptiva sobre los diferentes tramos, caracterizados por sus paradas terminantes, que componen cada una de las rutas de transporte ofrecidas en **Autobuses Interurbanos**.
Fuente	Portal de Datos Abiertos del CRTM.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero CSV (campos separados por ",").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	12 de septiembre de 2019.
Anomalías	- Presencia de campos completamente vacíos. - Valores incorrectos en el campo TIPOPARADA. - Valores nulos e incorrectos en el campo VELOCIDADTRAMOANTERIOR.
Decisiones	- Descartar campos vacíos. - Imputación manual de los valores incorrectos en TIPOPARADA. - Determinación de los valores nulos e incorrectos de VELOCIDADTRAMOANTERIOR a partir de la media calculada con el resto de los registros.

Figura B.28: Tabla Primera de M8 Tramos - Perfiles de Datos.

M8 TRAMOS		
Campo	**Descripción**	**Tipo de Datos**
CODIGOITINERARIO	Identificador de la ruta de transporte dentro de la red.	STRING
MODO	Modo de transporte en el que se transcurre la ruta.	STRING
NUMEROLINEAUSUARIO	Denominación de la línea a la que pertenece la ruta.	STRING
CODIGOSUBLINEA	Denominación de la sublínea a la que pertenece la ruta.	STRING
SENTIDO	Dirección de la ruta dentro de la línea.	STRING
NUMEROORDEN	Número de orden de la parada en la ruta.	INT
TIPOPARADA	Carácter correspondiente al tipo de la parada en la ruta.	STRING
LONGITUDTRAMOANTERIOR	Longitud, en metros, del tramo de ruta que termina en la parada.	FLOAT
VELOCIDADTRAMOANTERIOR	Velocidad media, en kilómetros por hora, a la que se recorre el tramo de ruta que termina en la parada.	FLOAT
CODIGOESTACION	Identificador de la estación dentro de la red de **Autobuses Interurbanos**.	STRING
Otros atributos: OBJECTID, IDTRAMO, FECHAACTUAL, CODIGOGESTIONLINEA, TIPOITINERARIO, CODIGOPOSTE, CODIGOANDEN, IDENTIFICADORTIPOPARADA, DENOMINACION, CODIGOPROVINCIA, CODIGOMUNICIPIO, MUNICIPIO, CORONATARIFARIA, DIRECCION, FECHAALTA, FECHAINICIO, FECHAFIN, MODOLINEA, MODOINTERCAMBIADOR, CODIGOINTERCAMBIADOR, CODPROV_LINEA, CODMUN_LINEA, IDFTRAMO, CODIGOOBSERVACION, DENOMINACION_SAE, IDFLINEA, SHAPE_Length.		

Figura B.29: Tabla Segunda de M8 Tramos - Perfiles de Datos.

M8 TRAMOS							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
CODIGOITINERARIO	1723	100%	308274	364183	-	-	[0-9]{6}
MODO	1	100%	8	8	-	-	8
NUMEROLINEAUSUARIO	336	100%	-	-	-	-	[0-9]{3}[ABCDEH]?(R1)?|N[0-9]{3}|VAC158
CODIGOSUBLINEA	1723	100%	-	-	-	-	N?[0-9]{3}[ABCDEH]?[0-9]{3}|VAC158[12]01
SENTIDO	2	100%	1	2	-	-	1|2
NUMEROORDEN	86	100%	1	86			[1-86]
TIPOPARADA	15	100%	-	-	-	-	C|I|T|IN|N|D|TN|CN|F|I2|ID|IX|S|TD|X
LONGITUDTRAMOANTERIOR	12290	100%	2,9439	46428,24	1439,38	2813,94	[0-9]{1,5},[0-9]+
VELOCIDADTRAMOANTERIOR	11	99,7%	-1	70	35,35	21,22	[0-9]{2},[0-9]+
CODIGOESTACION	6819	100%	06002	99987	-	-	[0-9]{1,5}
						Número Total de Registros	42660

Figura B.30: Tabla Tercera de M8 Tramos - Perfiles de Datos.

Datos de Petición al CRTM

TOPOLOGIA TRENES	
Definición	Contiene información topológica e identificativa de las estaciones de **Metro**, **Metro Ligero** y **Cercanías** de la red de transporte público de Madrid.
Fuente	Petición al Consorcio Regional de Transportes de Madrid.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero TXT (campos separados por "#").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	29 de febrero de 2020.
Anomalías	- Expresión de las coordenadas geográficas en los campos LATITUD y LONGITUD utilizando la coma (",") como separador decimal, lo cual es interpretado como cadena de caracteres. - Presencia de valores nulos en el campo DPAYPOINT.
Decisiones	- Sustitución del carácter "," por "." en los campos LATITUD y LONGITUD para la correcta interpretación de sus valores como de tipo numérico. - Imputación manual del máximo número de valores nulos en DPAYPOINT.

Figura B.31: Tabla Primera de Topología Trenes - Perfiles de Datos.

TOPOLOGIA TRENES		
Campo	**Descripción**	**Tipo de Datos**
IDFESTACION	Identificador de la estación dentro de la red de transporte público de Madrid.	STRING
LATITUD	Coordenada geográfica de latitud donde está ubicada la estación.	FLOAT(7,5)
LONGITUD	Coordenada geográfica de longitud donde está ubicada la estación.	FLOAT(6,5)
DPAYPOINT	Código identificativo de los terminales de registro de transacciones asociados a la estación.	STRING
Otros atributos: NUMODO, DENOMINACIONPARADA, CODIGOMUNICIPIO, CODIGOPOSTAL, CORONATARIFARIA, ZONIFICACION, XUTMETRS89, YUTMETRS89, XUTMED50, YUTMED50, IDACTOR, IDPARADA.		

Figura B.32: Tabla Segunda de Topología Trenes - Perfiles de Datos.

TOPOLOGIA TRENES							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
IDFESTACION	442	100%	-	-	-	-	(4|5|10)_[0-9]{3}
LATITUD	405	100%	40,03	40,82	40,42	0,09	40.[0-9]{4,5}
LONGITUD	423	100%	-4,27	-3,18	-3,71	0,12	-[34].[0-9]{4,5}
DPAYPOINT	458	65,3%	-	-	-	-	(02|04|A9|AD|AF|B1)_L[0-9]{1,2}_P[0-9]{1,3}
						Número Total de Registros	782

Figura B.33: Tabla Tercera de Topología Trenes - Perfiles de Datos.

TOPOLOGIA EMT	
Definición	Contiene información topológica e identificativa de las paradas de **Autobuses EMT** de la red de transporte público de Madrid.
Fuente	Petición al Consorcio Regional de Transportes de Madrid.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero TXT (campos separados por "#").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	29 de febrero de 2020.
Anomalías	- Expresión de las coordenadas geográficas en los campos LATITUD y LONGITUD utilizando la coma (",") como separador decimal, lo cual es interpretado como cadena de caracteres. - Presencia de valores nulos en el campo DPAYPOINT.
Decisiones	- Sustitución del carácter "," por "." en los campos LATITUD y LONGITUD para la correcta interpretación de sus valores como de tipo numérico. - Imputación manual del máximo número de valores nulos en DPAYPOINT.

Figura B.34: Tabla Primera de Topología EMT - Perfiles de Datos.

TOPOLOGIA EMT		
Campo	**Descripción**	**Tipo de Datos**
IDFPARADAGESTRA	Identificador de la parada dentro de la red de transporte público de Madrid.	STRING
IDPARADAE	Identificador de la parada dentro del código DPAYPOINT.	STRING
NULINGES	Código de la línea que transcurre por la estación.	STRING
NULINUSER	Denominación de la línea que transcurre por la estación.	STRING
LATITUD	Coordenada geográfica de latitud donde está ubicada la estación.	FLOAT(7,5)
LONGITUD	Coordenada geográfica de longitud donde está ubicada la estación.	FLOAT(6,5)
DPAYPOINT	Código identificativo de los terminales de registro de transacciones asociados a la estación.	STRING
Otros atributos: IDFEMPRESA, CONCESION, IDFLINEAGESTRA, DLINEA, DPARADA, CODIGOMUNICIPIO, CODIGOPOSTAL, CORONATARIFARIA, ZONIFICACION, XUTMETRS89, YUTMETRS89, XUTMED50, YUTMED50, IDACTOR, IDFLINEABIT, IDFPARADABIT, IDLINEA, IDPARADAI.		

Figura B.35: Tabla Segunda de Topología EMT - Perfiles de Datos.

TOPOLOGIA EMT							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
IDFPARADAGESTRA	4655	100%	6_10	6_6475	-	-	6_[0-9]{2,4}
IDPARADAE	4655	100%	1	6447	-	-	[0-9]{1,4}
NULINGES	206	100%	001	704	-	-	[0-9]{3}
NULINUSER	206	100%	-	-	-	-	SE7[0-9]{2}|[A-Z][0-9]{0,2}|[0-9]{1,3}(SF)?
LATITUD	3926	100%	40,33	40,52	40,42	0,04	40.[0-9]{4,5}
LONGITUD	4079	100%	-3,84	-3,57	-3,68	0,04	-[34].[0-9]{4,5}
DPAYPOINT	10430	96,2%	-	-	-	-	03_L[0-9]{1,3}_P[0-9]{1,4}
Número Total de Registros							10846

Figura B.36: Tabla Tercera de Topología EMT - Perfiles de Datos.

TOPOLOGIA INTERURBANOS	
Definición	Contiene información topológica e identificativa de las paradas de **Autobuses Interurbanos** de la red de transporte público de Madrid.
Fuente	Petición al Consorcio Regional de Transportes de Madrid.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero TXT (campos separados por "#").
Crecimiento	Conjunto de datos estático.
Fecha de Actualización	29 de febrero de 2020.
Anomalías	- Expresión de las coordenadas geográficas en los campos LATITUD y LONGITUD utilizando la coma (",") como separador decimal, lo cual es interpretado como cadena de caracteres. - Presencia de valores nulos en el campo DPAYPOINT. - Presencia de valores nulos en el campo IDLINEA.
Decisiones	- Sustitución del carácter "," por "." en los campos LATITUD y LONGITUD para la correcta interpretación de sus valores como de tipo numérico. - Imputación manual del máximo número de valores nulos en DPAYPOINT. - Mantener los valores nulos en el campo IDLINEA.

Figura B.37: Tabla Primera de Topología Interurbanos - Perfiles de Datos.

TOPOLOGIA INTERURBANOS		
Campo	**Descripción**	**Tipo de Datos**
IDFPARADAGESTRA	Identificador de la parada dentro de la red de transporte público de Madrid.	STRING
IDLINEA	Código de la línea que transcurre por la estación.	STRING
NULINUSER	Denominación de la línea que transcurre por la estación.	STRING
LATITUD	Coordenada geográfica de latitud donde está ubicada la estación.	FLOAT(7,5)
LONGITUD	Coordenada geográfica de longitud donde está ubicada la estación.	FLOAT(6,5)
DPAYPOINT	Código identificativo de los terminales de registro de transacciones asociados a la parada.	STRING
Otros atributos: IDFEMPRESA, CONCESION, IDFLINEAGESTRA, NULINGES, DLINEA, IDPARADAE, DPARADA, CODIGOMUNICIPIO, CODIGOPOSTAL, CORONATARIFARIA, ZONIFICACION, XUTMETRS89, YUTMETRS89, XUTMED50, YUTMED50, IDACTOR, IDFLINEABIT, IDFPARADABIT, IDPARADAI.		

Figura B.38: Tabla Segunda de Topología Interurbanos - Perfiles de Datos.

TOPOLOGIA INTERURBANOS							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
IDFPARADAGESTRA	8214	100%	8_06002	8_99987	-	-	8_[0-9]{5}
IDLINEA	367	95,2%	1	8722	-	-	[0-9]{1,4}
NULINUSER	355	100%	-	-	-	-	[0-9]{3}[ABCDEH]?(R1)?|CARGA|i1|N[0-9]{1,3}|Pi[12]|SE7|V5805|VAC-?[0-9]{3}
LATITUD	7537	100%	39,60	41,13	40,43	0,15	(39|40|41).[0-9]{4,5}
LONGITUD	7724	100%	-4,58	-2,82	-3,75	0,22	-[234].[0-9]{4,5}
DPAYPOINT	30483	95,2%	-	-	-	-	[0-9ABCDEF]{2}_L[0-9]{1,4}_P[0-9]{1,5}
Número Total de Registros							32830

Figura B.39: Tabla Tercera de Topología Interurbanos - Perfiles de Datos.

ENERO VIAJES 2019	
Definición	Contiene las transacciones de viaje realizadas en el mes de enero de 2019 en el transporte público de Madrid.
Fuente	Petición al Consorcio Regional de Transportes de Madrid.
Carga	Los datos han sido cargados en el directorio *data*.
Formato	Los datos están en un fichero TXT (campos separados por "#").
Crecimiento	Conjunto de datos dinámico.
Fecha de Actualización	31 de enero de 2019.
Anomalías	- Presencia de valores en el campo DPAYPOINT que no aparecen en los conjuntos de datos estáticos de la red de transporte público de Madrid.
Decisiones	- Descartar las transacciones con valores de DPAYPOINT no reconocibles, así como el resto de las transacciones realizadas con la misma tarjeta de pasajero en el día en cuestión.

Figura B.40: Tabla Primera de Enero Viajes 2019 - Perfiles de Datos.

ENERO VIAJES 2019		
Campo	**Descripción**	**Tipo de Datos**
TARJETA	Hash identificador de la tarjeta de transporte de un pasajero de la red de transporte público de Madrid.	STRING
FECHA	Fecha y hora correspondientes a la transacción de viaje.	DATETIME
TUSUARIO	Código asociado al perfil de usuario de la tarjeta de transporte.	STRING
TITULO	Código asociado al título de transporte del viaje.	STRING
DESCUENTO	Código de descuento aplicado al viaje.	STRING
DPAYPOINT	Código identificativo de los terminales de registro de transacciones asociados a la parada donde se ha realizado la transacción. Está formado por el código del operador, el número de la línea y el número de la parada.	STRING
IDTLV	Código indicador del modo de registro y la procedencia de la transacción.	STRING
CODVAL	Código de validación relativo al resultado y tipología de la transacción.	STRING
Otros atributos: -		

Figura B.41: Tabla Segunda de Enero Viajes 2019 - Perfiles de Datos.

ENERO VIAJES 2019							
Campo	Cardinalidad	Densidad	Rango				Formato
			MIN	MAX	AVG	STD	
TARJETA	4588825	100%	-	-	-	-	[A-Z0-9]{32}
FECHA	1212556	100%	01/01/2019 00:00:04	31/01/2019 23:59:58	-	-	2019-01-(0[1-9]|1[0-9]|2[0-9]|3[01]) (0[1-9]|1[0-9]|2[0-4]):[0-5][0-9]:[0-5][0-9]
TUSUARIO	10	100%	-	-	-	-	01|02|03|04|06|07|08|0B|0D|0F
TITULO	147	100%	-	-	-	-	[A-F0-9]{4}
DESCUENTO	6	99,85%	00	05	-	-	00|01|02|03|04|05
DPAYPOINT	31399	100%	-	-	-	-	[A-F0-9]{2}_L[0-9]{1,4}_P[0-9]{1,5}
IDTLV	4	100%	-	-	-	-	C0|C6|C1|G2
CODVAL	117	100%	0	177	-	-	[0-9]{1,3}
Número Total de Registros							151.048.520

Figura B.42: Tabla Tercera de Enero Viajes 2019 - Perfiles de Datos.

Apéndice C

Logical Datamap

En este apéndice se muestra el conjunto completo de tablas que constituyen el *logical datamap* planificado en el proyecto y que es descrito en el Apartado 4.3.3.

(1) M4 ESTACIONES **(2) M10 ESTACIONES** **(3) M5 ESTACIONES** **(4) M6 ESTACIONES** **(5) M8 ESTACIONES**	1.1. unionEstaciones	**CONJUNTO RESULTADO**
(1,2,3,4,5) IdEstacion	-	**ESTACIONES**
(1,2,3,4,5) Denominacion	-	**ESTACIONES**
(1,2,3,4,5) CodigoMunicipio	Renombrar a Municipio	**ESTACIONES**
(1,2,3,4,5) Distrito	-	**ESTACIONES**
(1,2,3,4,5) CoronaTarifaria	Renombrar a Corona	**ESTACIONES**
(1,2,3,4,5) X	-	**ESTACIONES**
(1,2,3,4,5) Y	-	**ESTACIONES**
Otros Atributos	✗	

Tabla C.1: Tabla Todas Estaciones - *Logical Datamap*.

(6) TOPOLOGIA TRENES **(7) TOPOLOGIA EMT** **(8) TOPOLOGIA INTERURBANOS**	1.1. unionTopologias	**CONJUNTO RESULTADO**
(6) IdfEstacion (7,8) IdfParadaGestra	Renombrar a IdUbicacion	**TOPOLOGIA**
(6,7,8) Latitud	-	**TOPOLOGIA**
(6,7,8) Longitud	-	**TOPOLOGIA**
(6,7,8) Dpaypoint	-	**TOPOLOGIA**
(6,7,8) NuLinUser	-	**TOPOLOGIA**
(6,7,8) NuLinGes	-	**TOPOLOGIA**
(6,7,8) IdParada	-	**TOPOLOGIA**
(6,7,8) IdLinea	-	**TOPOLOGIA**
Otros Atributos	✗	

Tabla C.2: Tabla Todas Topologías - *Logical Datamap.*

(9) M4 TRAMOS **(10) M10 TRAMOS** **(11) M5 TRAMOS** **(12) M6 TRAMOS** **(13) M8 TRAMOS**	1.1. unionTramos	**CONJUNTO RESULTADO**
(9,10,11,12,13) CodigoItinerario	-	**TRAMOS**
(9,10) IdfEstacion (11,12,13) Modo (11,12,13) CodigoEstacion	Renombrar a IdEstacion	**TRAMOS**
(9,10,11,12,13) Modo	-	**TRAMOS**
(9,10,11,12,13) NumeroLineaUsuario	-	**TRAMOS**
(9,10,11,12,13) Sentido	-	**TRAMOS**
(9,10,11,12,13) NumeroOrden	-	**TRAMOS**
(9,10,11,12,13) TipoParada	-	**TRAMOS**
(9,10,11,12,13) LongitudTramoAnterior	-	**TRAMOS**
(9,10,11,12,13) VelocidadTramoAnterior	-	**TRAMOS**
(13) CodigoSublinea	-	**TRAMOS**
Otros Atributos	✗	

Tabla C.3: Tabla Todos Tramos - *Logical Datamap.*

ESTACIONES	2.1 getUbicacion	2.2. obtenerParadasRutas	2.3. obtenerRutas	3.1. getDistanciaEstaciones	3.2. getDpaypoint	3.3. getTiempoParadas	4.1. tripChaining	5.1. agruparEtapas	**CONJUNTO RESULTADO**
IdEstacion	JOIN con TOPOLOGIA	-	-	RECORRIDO de ESTACIONES	-	-	-	-	**ESTACIONES y DISTANCIA ESTACIONES**
Denominacion	-	-	-	-	-	-	-	-	**ESTACIONES**
CodigoMunicipio	-	-	-	-	-	-	-	-	**ESTACIONES**
Distrito	-	-	-	-	-	-	-	-	**ESTACIONES**
CoronaTarifaria	-	-	-	-	-	-	-	-	**ESTACIONES**
X	PROCESO	✗							
Y	PROCESO	✗							
	Creación de Latitud	-	-	PROCESO	-	-	-	-	**ESTACIONES**
	Creación de Longitud	-	-	PROCESO	-	-	-	-	**ESTACIONES**
				Creación de Distancia	-	-	-	-	**ESTACIONES y DISTANCIA ESTACIONES**

Tabla C.4: Tabla Estaciones - *Logical Datamap.*

TOPOLOGÍA	2.1 getUbicacion	2.2. obtenerParadasRutas	2.3. obtenerRutas	3.1. getDistanciaEstaciones	3.2. getDpaypoint	3.3. getTiempoParadas	4.1. tripChaining	5.1. agruparEtapas	**CONJUNTO RESULTADO**
IdUbicacion	JOIN con ESTACIONES	-	-	-	JOIN con PARADAS RUTAS	-	-	-	-
Latitud	PROCESO	-	-	-	-	-	-	-	-
Longitud	PROCESO	-	-	-	-	-	-	-	-
Dpaypoint	-	-	-	-	PROCESO	-	-	-	-
NuLinUser	-	-	-	-	PROCESO	-	-	-	-
NuLinGes	-	-	-	-	PROCESO	-	-	-	-
IdParada	-	-	-	-	PROCESO	-	-	-	-
IdLinea	-	-	-	-	PROCESO	-	-	-	-

Tabla C.5: Tabla Topología - *Logical Datamap.*

TRAMOS	2.1 getUbicacion	2.2. obtenerParadasRutas	2.3. obtenerRutas	3.1. getDistanciaEstaciones	3.2. getDpaypoint	3.3. getTiempoParadas	4.1. tripChaining	5.1. agruparEtapas	**CONJUNTO RESULTADO**
CodigoItinerario	-	Creación de PARADAS RUTAS	Creación de RUTAS	-	-	-	-	-	-
Modo	-	-	Creación de RUTAS	-	-	-	-	-	-
IdEstacion	-	Creación de PARADAS RUTAS	-	-	-	-	-	-	-
NumeroLineaUsuario	-	-	-	-	-	-	-	-	-
Sentido	-	-	Creación de RUTAS	-	-	-	-	-	-
NumeroOrden	-	Creación de PARADAS RUTAS	Creación de RUTAS	-	-	-	-	-	-
TipoParada	-	Creación de PARADAS RUTAS	-	-	-	-	-	-	-
LongitudTramoAnterior	-	Creación de PARADAS RUTAS	-	PROCESO	-	-	-	-	-
VelocidadTramoAnterior	-	Creación de PARADAS RUTAS	-	PROCESO	-	-	-	-	-
CodigoSublinea	-	-	Creación de RUTAS	-	-	-	-	-	-

Tabla C.6: Tabla Tramos - *Logical Datamap.*

PARADAS RUTAS	3.1. getDistanciaEstaciones	3.2. getDpaypoint	3.3. getTiempoParadas	4.1. tripChaining	5.1. agruparEtapas	**CONJUNTO RESULTADO**
IdParadaRuta	-	-	RECORRIDO de PARADAS RUTAS	-	-	**PARADAS RUTAS y TIEMPO PARADAS**
IdEstacion	-	JOIN con TOPOLOGÍA	-	-	-	**PARADAS RUTAS**
CodigoItinerario	-	JOIN con RUTAS	JOIN con RUTAS	-	-	**PARADAS RUTAS y TIEMPO PARADAS**
NumeroOrden	-	-	PROCESO	-	-	**PARADAS RUTAS**
TipoParada	-	-	PROCESO	-	-	**PARADAS RUTAS**
LongitudTramoAnterior	-	-	PROCESO	-	-	-
ValocidadTramoAnterior	-	-	PROCESO	-	-	-
		Creación de Dpaypoint	-	JOIN con ENERO VIAJES 2019	-	**PARADAS RUTAS**
			Creación de Tiempo	-	-	**TIEMPO PARADAS**

Tabla C.7: Tabla Paradas Rutas - *Logical Datamap.*

RUTAS	3.1. getDistanciaEstaciones	3.2. getDpaypoint	3.3. getTiempoParadas	4.1. tripChaining	5.1. agruparEtapas	**CONJUNTO RESULTADO**
CodigoItinerario	-	JOIN con PARADAS RUTAS	JOIN con PARADAS RUTAS	-	-	**RUTAS**
Modo	-	-	PROCESO	-	-	**RUTAS**
NumeroLineaUsuario	-	-	PROCESO	-	-	**RUTAS**
Sentido	-	-	PROCESO	-	-	**RUTAS**
CodigoSublinea	-	-	PROCESO	-	-	**RUTAS**

Tabla C.8: Tabla Rutas - *Logical Datamap.*

(14) ENERO VIAJES 2019	2.1. preprocesar-Transacciones	4.1. tripChaining	5.1. agruparEtapas	**CONJUNTO RESULTADO**
IdEtapa	PROCESO	-	-	**ETAPAS**
Tarjeta	PROCESO	PROCESO	PROCESO	**VIAJES**
Fecha	PROCESO	PROCESO	PROCESO	**ETAPAS y VIAJES**
TUsuario	PROCESO	✗		
Titulo	PROCESO	✗		
Descuento	PROCESO	✗		
Dpaypoint	PROCESO	JOIN con PARADAS RUTAS	-	-
IdTlv	PROCESO	✗		
CodVal	PROCESO	PROCESO	-	-
		Creación de ParadaSubida	-	**ETAPAS**
		Creación de ParadaBajada	-	**ETAPAS**
		Creación de FechaHoraFin	-	**ETAPAS y VIAJES**
			Creación de IdViaje	**ETAPAS y VIAJES**

Tabla C.9: Tabla Enero Viajes 2019 - *logical datamap*.

Apéndice D

Herramientas Utilizadas

Este apéndice comprende la descripción de las diferentes herramientas empleadas en el proyecto junto con las librerías de Python instaladas para su implementación. Asimismo, se muestra una tabla con las versiones concretas de las librerías de Python instaladas.

A continuación, se presentan cada una de las herramientas y utilidades consideradas en el desarrollo de este proyecto, agrupadas en distintas categorías para su mejor organización.

Entorno de Desarrollo.

- **Máquina Virtual Linux**: el desarrollo del proyecto se ha llevado a cabo sobre una máquina virtual cedida por la Universidad de Valladolid. Dicha máquina virtual tiene instalado el sistema operativo Ubuntu Desktop 20.04, una distribución de Linux.
- **Anaconda**: se trata de una distribución libre de Python especialmente indicada para trabajar en proyectos de ciencia de datos y aprendizaje automático. Integra numerosas herramientas dispuestas para distintas finalidades y facilita la instalación directa de librerías y la gestión de entornos de desarrollo virtuales.
- **Jupyter Notebooks**: es una herramienta que permite crear y compartir documentos de código junto con otros elementos, tales como texto, imágenes y ecuaciones, con el objetivo de enriquecer su presentación y favorecer su accesibilidad a otros desarrolladores. En lo relativo al proyecto, todo el código de implementación se ha ido desarrollando a través de distintos cuadernos de Jupyter Notebook. Esta herramienta ya aparece instalada en la distribución Anaconda utilizada.

Librerías de Python

- **Pip**: es un sistema de gestión de paquetes utilizado para administración e instalación de librerías de Python junto con la gestión de sus dependencias. Es la librería base por medio de la cual se instalan el resto de los paquetes necesarios en el proyecto.

- **Pandas**: es una librería destinada para llevar a cabo el análisis y manipulación de datos sobre diferentes tipos de estructuras. En este proyecto, todos los conjuntos de datos utilizados son cargados como *Dataframes*.
- **Pandas Profiling**: es una herramienta para generar informes de perfiles de datos a partir de *Dataframes* de Pandas. En este proyecto se han empleado los informes producidos por Pandas Profiling para realizar una exploración exhaustiva de los conjuntos de datos disponibles. De hecho, parte de los datos especificados en los perfiles de datos, presentes en el Apéndice B, han sido derivados de los resultados proporcionados por esta herramienta.
- **Matplotlib**: es una biblioteca de visualización de datos para generar gráficos de tipos muy variados a partir de datos contenidos en listas y arrays de Python o en estructuras de Numpy. En concreto, en este proyecto se utiliza esta librería para implementar los gráficos de frecuencia de transacciones por hora y para visualizar la corrección de los grafos diseñados para modelar las redes de metro y cercanías de Madrid.
- **Seaborn**: es una librería de visualización de datos basada en Matplotlib que proporciona una interfaz de alto nivel para facilitar el diseño de gráficos de visualización, permitiendo su implementación con un menor número de líneas de código. En este proyecto se ha utilizado para implementar los mapas de calor con los que se representan las diferentes matrices OD de tránsito diseñadas.
- **Plotly**: es una librería de visualización de datos para elaborar gráficos interactivos de calidad de una forma sencilla. La decisión de utilizar esta librería en el proyecto se debe a que dispone de funciones básicas para la visualización de datos geográficos sobre mapas. En concreto, todos los mapas diseñados para visualizar resultados relativos a los viajes realizados en la red de transporte de Madrid han sido implementados con Plotly.
- **Numpy**: es una librería de cálculo numérico especializada en la manipulación de vectores, matrices y otras estructuras de datos de tipo numérico. En este proyecto, su uso se ha limitado a la realización de operaciones de reescalado de datos de cara a su disposición en ciertos gráficos.
- **Networkx**: es una librería enfocada al estudio y análisis de redes complejas y a su modelización por medio de grafos. En este proyecto ha sido utilizada para modelar la topología de las redes de metro y cercanías de Madrid.
- **Pyproj**: es una librería para la realización de operaciones de proyección de datos geográficos y su conversión entre distintos sistemas de referencia. En este proyecto se utiliza para transformar ciertas coordenadas geográficas al sistema de referencia geográfico correspondiente al estándar GPS.

- **Geopy**: es una librería para la geolocalización de ubicaciones expresadas en múltiples formatos y el cálculo de diferentes funciones entre coordenadas geográficas. En este proyecto su uso se reduce al cálculo de la distancia ortodrómica entre paradas de la red de transporte de Madrid.

Exploración de Datos.

- **OpenRefine**: es una aplicación de escritorio de código abierto destinada a la exploración, limpieza y transformación de conjuntos de datos. En este proyecto se ha utilizado para complementar la tarea exploración de datos, también realizada en Python.

Modelado de Diagramas.

- **Draw.io**: es una aplicación web de código abierto orientada al diseño gráfico y la elaboración de diferentes tipos de diagramas. Cuenta con multitud de figuras y componentes editables por el usuario para crear diagramas personalizados y adaptados a sus necesidades. En este proyecto se ha utilizado para la creación de diagramas como el *pipeline* de transformación de datos y para la adaptación de ciertos gráficos y mapas, con el objetivo de enriquecer la información mostrada.
- **Tom's Planner**: es una aplicación web de pago, aunque con versión gratuita para uso personal, específicamente designada para la creación de diagramas de Gantt con los que diseñar la planificación temporal de un proyecto. Evidentemente, en este proyecto se ha empleado para la creación del diagrama de Gantt de planificación temporal de las Historias de Usuario a completar a lo largo de los sprints comprendidos en el proyecto.
- **Tom's Planner**: es una aplicación web de pago, aunque con versión gratuita para uso personal, específicamente designada para la creación de diagramas de Gantt con los que diseñar la planificación temporal de un proyecto. Evidentemente, en este proyecto se ha empleado para la creación del diagrama de Gantt de planificación temporal de los sprints comprendidos en el proyecto.
- **Creately**: es una aplicación web de pago, aunque con un plan básico gratuito, para la creación de diagramas de todo tipo y que facilita un uso colaborativo por parte de equipos de trabajo. En este proyecto se ha utilizado únicamente para crear el modelo lógico de datos.

Redacción de Memoria.

- **TeXstudio**: la redacción del presente documento se ha realizado a través de TeXstudio, un editor de código abierto de LaTeX destinado a la creación de documentación

de alta calidad. A diferencia de los editores de texto de tipo WYSIWYG (*What You See Is What You Get*), la edición de texto en el lenguaje LaTeX se basa en una clara separación entre contenido y formato, encargándose el motor de LaTeX de este último aspecto.

Por último, en la Figura D.1 se muestran las versiones concretas instaladas de las librerías de Python utilizadas en el proyecto. Un aspecto a tener en cuenta es que dicha instalación se ha realizado sobre un entorno virtual con Python 3.8, siguiendo las buenas prácticas de gestión de dependencias de paquetes y aislamiento de proyectos que evite el conflicto de versiones.

Librería	**Versión Instalada**
pip	22.1.2
pandas	1.4.3
pandas_profiling	3.2.0
matplotlib	3.5.2
seaborn	0.11.2
plotly	5.9.0
numpy	1.23.1
networkx	2.8.4
pyproj	3.3.1
geopy	2.2.0

Tabla D.1: Versiones de las librerías de Python instaladas.

Bibliografía

[1] Juan Carlos Tovar. The 42 v's of big data and data science. https://forum.huawei.com/enterprise/es/las-5vs-del-big-data/thread/846137-100759, 2022. Accedido: 27-06-2022.

[2] Jim Frazer. The nine critical applications of a smart city. https://www.arcweb.com/blog/nine-critical-applications-smart-city, 2018. Accedido: 28-06-2022.

[3] Antonio Nunes, Teresa Dias, and João Falcão e Cunha. Passenger journey destination estimation from automated fare collection system data using spatial validation. *IEEE Transactions on Intelligent Transportation Systems*, 17:1–10, 08 2015.

[4] Mohammed Mohammed and Jimi Oke. Origin-destination inference in public transportation systems: A comprehensive review. *International Journal of Transportation Science and Technology*, 2022.

[5] Etikaf Hussain, Ashish Bhaskar, and Edward Chung. Transit od matrix estimation using smartcard data: Recent developments and future research challenges. *Transportation Research Part C Emerging Technologies*, 125, 04 2021.

[6] Behrang Assemi, Azalden Alsger, Mahboobeh Moghaddam, Mark Hickman, and Mahmoud Mesbah. Improving alighting stop inference accuracy in the trip chaining method using neural networks. *Public Transport*, 12:89–121, 03 2020.

[7] Zhanhong Cheng, Martin Trépanier, and Lijun Sun. Probabilistic model for destination inference and travel pattern mining from smart card data. *Transportation*, 48, 08 2021.

[8] Tian Li, Dazhi Sun, Jing Peng, and Kaixi Yang. Smart card data mining of public transport destination: A literature review. *Information*, 9:18, 01 2018.

[9] Consorcio Regional de Transportes de Madrid. 0c. plano esquemático de la red de metro, metro ligero y cercanías en el ámbito de toda la comunidad de madrid. https://www.crtm.es/atencion-al-cliente/area-de-descargas/planos/

serie-0/0c-plano-metro-y-cercanias-comunidad-madrid.aspx, 2022. Accedido: 02-07-2022.

[10] Consorcio Regional de Transportes de Madrid. Zonas tarifarias de la red de transporte público de madrid. https://www.crtm.es/billetes-y-tarifas/zonas-tarifarias.aspx, 2022. Accedido: 02-07-2022.

[11] Consorcio Regional de Transportes de Madrid. Legislación y normativa. consorcio regional de transportes de madrid. https://transparencia.crtm.es/legislacion-y-normativa.aspx, 2022. Accedido: 02-07-2022.

[12] IBM. Conceptos básicos de ayuda de crisp-dm. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview, 2021. Accedido: 03-08-2022.

[13] Miguel A. Martínez-Prieto. *Apuntes de la asignatura Arquitecturas Big Data*. Máster en Inteligencia de Negocio y Big Data en Entornos Seguros, 2021.

[14] Senem Sagiroglu and Duygu Sinanc. Big data: A review. In *Big data: A review*, pages 42–47, 05 2013.

[15] Tom Shafer. The 42 v's of big data and data science. https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html, 2017. Accedido: 27-06-2022.

[16] Rebecca Hammons and Joel Myers. Architects of our future: Redefining smart cities to be people-centric and socially responsible. *IEEE Internet of Things Magazine*, 2:10–14, 06 2019.

[17] Al Nuaimi, E. Al Neyadi, H. Mohamed, N. et al. Applications of big data to smart cities. *Journal of Internet Services and Applications*, 6:25, 12 2015.

[18] Soledad Pellicer, Guadalupe Santa, Andres L. Bleda, Rafael Maestre, Antonio J. Jara, and Antonio Gomez Skarmeta. A global perspective of smart cities: A survey. In *2013 Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, pages 439–444, 2013.

[19] James Barry, Robert Newhouser, Adam Rahbee, and Shermeen Sayeda. Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record*, 1817:183–187, 01 2002.

[20] Firas Alrawi. The importance of intelligent transport systems in the preservation of the environment and reduction of harmful gases. *Transportation Research Procedia*, 24:197–203, 2017. 3rd Conference on Sustainable Urban Mobility, 3rd CSUM 2016, 26 – 27 May 2016, Volos, Greece.

[21] Steve Mazur. An introduction to smart transportation: Benefits and examples. https://www.digi.com/blog/post/introduction-to-smart-transportation-benefits, 2020. Accedido: 29-06-2022.

[22] Neema Nassir, Mark Hickman, and Zhenliang Ma. Activity detection and transfer identification for public transit fare card data. *Transportation*, 42:1–23, 07 2015.

[23] Azalden Alsger, Behrang Assemi, Mahmoud Mesbah, and Luís Ferreira. Validating and improving public transport origin–destination estimation algorithm using smart card fare data. *Transportation Research Part C-emerging Technologies*, 68:490–506, 2016.

[24] Google. Descripción general gtfs. https://developers.google.com/transit/gtfs, 2021. Accedido: 30-06-2022.

[25] Google. Referencia gtfs schedule. https://www.digi.com/blog/post/introduction-to-smart-transportation-benefits, 2022. Accedido: 30-06-2022.

[26] Jinhua Zhao, Adam Rahbee, and Nigel Wilson. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Comp.-Aided Civil and Infrastruct. Engineering*, 22:376–387, 07 2007.

[27] Janine M. Farzin. Constructing an automated bus origin–destination matrix using farecard and global positioning system data in são paulo, brazil. *Transportation Research Record*, 2072(1):30–37, 2008.

[28] Catherine Seaborn, John Attanucci, and Nigel H. M. Wilson. Analyzing multimodal public transport journeys in london with smart card fare payment data. *Transportation Research Record*, 2121(1):55–62, 2009.

[29] Wei Wang, John Attanucci, and Nigel Wilson. Bus passenger origin-destination estimation and related analyses using automated data collection systems. *Journal of Public Transportation*, 14:131–150, 12 2011.

[30] Marcela Munizaga and Carolina Palma. Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smartcard data from santiago, chile. *Transportation Research Part C-Emerging Technologies*, 24:9–18, 10 2012.

[31] Jason B. Gordon, Harilaos N. Koutsopoulos, Nigel H. M. Wilson, and John P. Attanucci. Automated inference of linked transit journeys in london using fare-transaction and vehicle location data. *Transportation Research Record*, 2343(1):17–24, 2013.

[32] Pramesh Kumar, Alireza Khani, and Qing He. A robust method for estimating transit passenger trajectories using automated data. *Transportation Research Part C: Emerging Technologies*, 95:731–747, 2018.

[33] Fenfan Yan, Chao Yang, and Satish Ukkusuri. Alighting stop determination using two-step algorithms in bus transit systems. *Transportmetrica A: Transport Science*, 15:1–29, 05 2019.

[34] Di Huang, Jun Yu, Shiyu Shen, Zhekang Li, Luyun Zhao, and Cheng Gong. A method for bus od matrix estimation using multisource data. *Journal of Advanced Transportation*, 2020:1–13, 03 2020.

[35] Boletín Oficial del Estado. Ley 5/1985, de 16 de mayo, de creación del consorcio regional de transportes públicos regulares de madrid. https://www.boe.es/eli/es-md/l/1985/05/16/5, 1985. Accedido: 01-07-2022.

[36] Consorcio Regional de Transportes de Madrid. Conócenos. consorcio regional de transportes de madrid. https://www.crtm.es/conocenos.aspx, 2022. Accedido: 02-07-2022.

[37] Consorcio Regional de Transportes de Madrid. Tu transporte público. consorcio regional de transportes de madrid. https://www.crtm.es/tu-transporte-publico.aspx, 2022. Accedido: 02-07-2022.

[38] Consorcio Regional de Transportes de Madrid. Datos abiertos crtm. https://data-crtm.opendata.arcgis.com/, 2022. Accedido: 03-07-2022.

[39] Martin Trépanier, Nicolas Tranchant, and Robert Chapleau. Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1):1–14, 2007.

[40] James J. Barry, Robert Freimer, and Howard Slavin. Use of entry-only automatic fare collection data to estimate linked transit trips in new york city. *Transportation Research Record*, 2112(1):53–61, 2009.

[41] Neema Nassir, Alireza Khani, Sang Gu Lee, Hyunsoo Noh, and Mark Hickman. Transit stop-level origin–destination estimation through use of transit schedule and automated data collection system. *Transportation Research Record*, 2263(1):140–150, 2011.

[42] Chen Jun and Yang Dongyuan. Estimating smart card commuters origin-destination distribution based on apts data. *Journal of Transportation Systems Engineering and Information Technology*, 13:47–53, 08 2013.

[43] Jaeyoung Jung and Keemin Sohn. Deep learning architecture to forecast destinations of bus passengers from entry-only smart-card data. *IET Intelligent Transport Systems*, 11:1. 334–339, 03 2017.

[44] Inmook Lee, Shin-Hyung Cho, Kyoungtae Kim, Seung-Young Kho, and Dong-Kyu Kim. Travel pattern-based bus trip origin-destination estimation using smart card data. *PLOS ONE*, 17(6):1–20, 06 2022.

[45] Miguel A. Martínez-Prieto, Jorge Silvestre, Anibal Bregon, and José Ignacio Farrán. Hacia la consolidación de las aulas Ágiles. *Actas de las XXVI Jornadas sobre Enseñanza Universitaria de la Informática*, pages 29–36, 2020.

[46] Ken Schwaber and Jeff Sutherland. *La Guía Scrum*. Scrum.org, 2020.

[47] Peter Chapman, Janet Clinton, Randy Kerber, Tom Khabaza, Thomas P. Reinartz, Colin Shearer, and Richard Wirth. Crisp-dm 1.0: Step-by-step data mining guide. In *CRISP-DM 1.0: Step-by-step data mining guide*, 2000.

[48] Mountain Goat Software. Planning poker. https://www.mountaingoatsoftware.com/agile/planning-poker, 2022. Accedido: 16-08-2022.

[49] GISGeography. World geodetic system (wgs84). https://gisgeography.com/wgs84-world-geodetic-system/, 2022. Accedido: 12-08-2022.

[50] Diego Alonso. Qué son los códigos epsg / srid y su vinculación con postgis. https://www.crtm.es/billetes-y-tarifas/zonas-tarifarias.aspx, 2022. Accedido: 12-08-2022.

[51] Nur Hanis Kasehyani. Evaluation of pedestrian walking speed in rail transit terminal. *International Journal of Integrated Engineering*, 11(9):026–036, Dec. 2019.

[52] Marcela Munizaga, Flavio Devillaine, Claudio Navarrete, and Diego Silva. Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies*, 44:70–79, 2014.

[53] Azalden Alsger, Behrang Assemi, Mahmoud Mesbah, and Luis Ferreira. Validating and improving public transport origin–destination estimation algorithm using smart card fare data. *Transportation Research Part C Emerging Technologies*, 68:490–506, 07 2016.

[54] Daniel Álvarez Gil. Conceptos básicos de ayuda de crisp-dm. https://www.adictosaltrabajo.com/2021/01/14/metodologia-crisp-dm/, 2021. Accedido: 04-08-2022.