Facultad de Ciencias Económicas y Empresariales, ICADE

APLICABILIDAD DE HERRAMIENTAS DE BIG DATA Y BUSINESS ANALYTICS A LA MEJORA DE LA ECOMOVILIDAD URBANA: EL CASO BICIMAD

Clave: 201600330

MADRID | Junio 2020

Resumen:

La ciencia de los datos es un campo relativamente nuevo, que nos permite analizar los procesos de cualquier empresa, institución u órgano gubernamental y mejorar su eficiencia. Con la aglomeración de la población en las ciudades, nuevos medios de movilidad urbana han surgido y la mayoría de ellos se apoyan, estratégicamente, en datos y el análisis de estos, ofreciendo a sus usuarios una mejor experiencia de uso/servicio/venta etc.

Debido a tal combinación de eventos, este documento intenta de una manera genérica, acercar a cualquier lector a este nuevo campo a través de una evolución histórica del "Big Data" nombrando algunas de las herramientas más utilizadas hoy en día y de una manera específica, demostrar mediante el análisis de datos cómo podemos solventar algunos de los problemas en el sector de la movilidad urbana, identificando iniciativas actuales que se apoyan en el uso de los datos y elaborando un extenso análisis sobre el sistema de bicicletas eléctricas por alquiler de la ciudad de Madrid, denominado BiciMAD.

Palabras clave:

Ciencia de los Datos, Datos Masivos, Sostenibilidad, Análisis de Negocios, Análisis de Datos, Movilidad Urbana, Transporte, Transformación Digital, Aplicabilidad de Herramientas, Inteligencia Empresarial, Clústeres, Bicicletas.

Abstract:

Data science is a relatively new field, which allows us to analyze the processes of any company, institution or government body and improve their efficiency. With the agglomeration of the population in cities, new means of urban mobility have emerged and most of them are strategically supported by data and its analysis, offering their users a better experience of use/service/sale etc.

Due to such a combination of events, this document tries in a generic way, to bring any reader closer to this new field through a vision and historical evolution of Big Data naming some of the most used tools nowadays and in a specific way, to demonstrate through data analysis how we can solve some of the problems in the urban mobility sector, identifying current initiatives that rely on the use of the data and preparing an extensive analysis of the city of Madrid's electric bicycle rental system, called BiciMAD.

Key words:

Data Science, Big Data, Sustainability, Business Analytics, Data Analysis, Urban Mobility, Transport, Digital Transformation, Tool Applicability, Business Intelligence, Clustering, Bicycles.

A mi tutora, [redacted] *, por su paciencia, dinamismo y consejo a lo largo del proceso*

A mi familia por apoyarme y escucharme en todo momento ante situaciones adversas

A [redacted] *, por introducirme al análisis de datos en los negocios*

Y a [redacted] *, por introducirme a las técnicas de aprendizaje no supervisado*

Índice

Ilustraciones/Figuras*

*las ilustraciones de los anexos no se incluyen en la tabla de figuras.

Tablas

1. Introducción y Motivación del trabajo

1.1 Estado de la cuestión, justificación y motivación

El sector de la movilidad urbana cobra vital importancia en las ciudades. Problemas de carácter medioambiental o la superpoblación han obligado a las ciudades a modernizarse con alternativas, más eficientes y limpias. Estos hechos dan cabida a nuevos modelos de transporte, como los vehículos compartidos, patinetes, motocicletas y bicicletas impulsados por fuentes alternativas de energía, pero también dan cabida a las métricas y a la ciencia de los datos para la mejora de la eficiencia de sus sistemas y, por consiguiente, de las ciudades. Se debe recalcar que el valor real de los datos recolectados en bruto solo puede ser comprendido una vez transformados en información. La arcilla, como los datos en bruto, no tiene un valor material alto antes de ser transformado. No obstante, muchos de los edificios en Londres utilizan ladrillos de arcilla amarillenta guardando un gran valor percibido por la sociedad (McQueen, 2017).

Si alguna vez ha utilizado algún medio de transporte, tanto público como privado, se habrá dado cuenta de que la estadística es de importancia mayúscula para el buen funcionamiento del sistema de transporte de una ciudad. Algunos ejemplos del uso de estadística en la movilidad urbana puede ser el número de minutos para llegar a una parada en un autobús de la Empresa Municipal de Transporte (EMT) de Madrid, el porcentaje de batería restante de una bicicleta eléctrica, el nivel de demanda en Uber o los precios preestablecidos por tarifa fijada en un taxi.

Un profesor de la universidad de Northeastern y director de analítica avanzada en L.E.K Consulting, en un informe sobre el poder de la estadística y del uso del “Big Data”, transmite el siguiente mensaje (Breur, 2016):

> *“No analizamos los datos para encontrar hechos "agradables de conocer". Cuando se presentan hechos de negocio notables, el objetivo es estimular nuevas ideas e innovar las prácticas de negocio existentes. Cuando las empresas intentan actuar sobre estos hallazgos analíticos, lo hacen (casi) siempre de forma experimental y poco sistemática. La implementación es donde la teoría se pone a prueba”* (p. 61).

Para garantizar la eficiencia logística de los sistemas de transporte y dar un mejor servicio al usuario es necesario recolectar, tratar y procesar una gran cantidad de datos, con el objetivo de transformar estos datos en información relevante. Esta información puede ser utilizada para mejorar la eficiencia de los procesos, la oferta al cliente y, en general, para tomar mejores decisiones de negocio. A modo de ejemplo, una reciente alianza ha sido la de Moovit, la aplicación de MaaS o "Mobility as a Service", y la Empresa Municipal de Transporte de Madrid de Madrid (EMT) para utilizar los datos generados por la aplicación y descifrar los ámbitos de los usuarios en el uso del transporte (Ayuntamiento de Madrid, 2020). De la misma manera, la revista McKinsey Quarterly, publicó en febrero de 2019 un artículo sobre el segundo gran punto de inflexión de la movilidad y el transporte, después de mostrar el primero con el paso del caballo a los motores de vapor y combustión, y de cómo este se centra en la recolección y uso de datos por el vehículo, combatiendo problemas como el exceso de tráfico, los accidentes o la contaminación (Dhawan, Hensley, Padhi, & Tschiesner, 2019).

Por todo esto, nos enfrentamos a una industria en constante evolución, que ha crecido exponencialmente, y donde se verán grandes cambios en los próximos diez años gracias al "Big Data" y el análisis de datos. La tecnología es más accesible que nunca y por ello, resulta interesante realizar este trabajo con el fin de conocer estos nuevos campos de investigación, ver qué iniciativas se llevan a cabo en el sector de la movilidad urbana y demostrar, mediante el caso práctico de BiciMAD (empresa de transporte público a cargo de la gestión y alquiler de las bicicletas eléctricas para la ciudad de Madrid), como el uso de los datos ayuda a las empresas a tomar mejores decisiones de negocio.

Concretamente, se ha elegido BiciMAD dado que es un modelo de transporte novedoso, útil y sostenible con el medioambiente. Según el índice urbano de movilidad por HERE Technologies, Madrid se sitúa en el primer puesto por el porcentaje de zonas verdes en la ciudad (de 38 ciudades analizadas). Sin embargo, Madrid baja hasta el vigesimocuarto puesto en el número de bicicletas ancladas en una estación por cada mil habitantes, lo que sugiere problemas a nivel operacional o poco conocimiento de este sistema en Madrid, una razón más por la cual se realiza este trabajo (HERE, 2020).

1.2 Objetivos

El objetivo general de este trabajo es demostrar el potencial que tienen las nuevas tecnologías, el "Big Data" y el análisis de datos para dar respuesta a los problemas que surgen dentro del ámbito de la movilidad urbana y mejorar la eficiencia de los sistemas de transporte dentro de las ciudades. Además, no sólo se pretende abordar este objetivo desde una perspectiva teórica y descriptiva, sino también desde una perspectiva más práctica, mediante la elaboración de un caso práctico con datos reales del sistema BiciMAD.

Para conseguir este objetivo general será necesario abordar los siguientes objetivos específicos, los cuales son:

1. Estudiar los principales conceptos relacionados con el "Big Data" y la analítica de datos, identificando además las principales herramientas y tecnologías utilizadas por las empresas para almacenar, procesar, analizar y visualizar los datos.
2. Ahondar en los problemas que surgen en el ámbito de la movilidad urbana, así como las soluciones que se están dando actualmente utilizando las nuevas tecnologías, el "Big Data" y el análisis de datos.
3. Elaborar un caso práctico sobre el sistema BiciMAD. Con este caso se pretende demostrar cómo la utilización de diferentes herramientas de análisis de datos contribuye a un mejor conocimiento de negocio y de esta manera ayuden a tomar mejores decisiones.

En cuanto al caso práctico de BiciMAD, cabe aclarar que, gran parte de este se desarrolla en torno a un importante problema operacional/logístico que surge en los sistemas de transporte urbano en bicicleta: la necesidad de balancear el sistema. Debido a que hay estaciones que reciben más bicicletas que las que salen de ellas (y viceversa), se requiere que el sistema sea constantemente balanceado moviendo bicicletas desde donde hay un exceso de ellas a donde hay un defecto, y garantizar así la disponibilidad de bicicletas cuando el cliente lo requiere. En el caso se muestra cómo las técnicas de análisis de datos pueden ayudar a corregir este problema y balancear el sistema de una forma eficiente.

1.3 Metodología

Para cumplir con el primer objetivo específico mencionado anteriormente, se explica la evolución de las principales tecnologías "Big Data" y de cómo las empresas se han adaptado a esto a través de la computación paralela y herramientas de "Business Analytics". Por otro lado, para cumplir con el segundo objetivo específico, se entra en materia con un estudio del sector transporte urbano, para comprender los grandes avances en temas clave como el tráfico, la sostenibilidad, la seguridad vial o el vehículo autónomo. Para la consecución de estos objetivos, se consultarán varias fuentes académicas, artículos periodísticos y de divulgación especializados en "Data Science" y libros escritos sobre la aplicación del "Big Data" al sector de la movilidad urbana, siempre con la intención de responder a los objetivos de este estudio.

En cuanto al tercer objetivo específico, con el desarrollo del caso práctico, se han obtenido los datos de uso por los usuarios y el estado de las estaciones de BiciMAD del portal OPENDATA de la EMT de Madrid, donde ha sido necesario realizar un trabajo de correcta lectura de los datos y un preprocesamiento de estos, con el fin de que se pueda desempeñar un correcto análisis. A través de un análisis exploratorio de los datos y con la utilización de diferentes herramientas de visualización, se demuestra cómo BiciMAD puede dar respuesta a preguntas del tipo: ¿en qué rango de edad se sitúan los usuarios más frecuentes?, ¿qué tipo de usuarios realizan los trayectos de mayor duración? Para ello, además de utilizar el análisis exploratorio y las herramientas de visualización, también se realizan diferentes análisis clúster con el objeto de dar respuesta a las preguntas: ¿las estaciones se comportan de manera diferente respecto a su ocupación, capacidad o número de reservas?, ¿cómo podemos balancear y gestionar el sistema de una manera más eficiente? demostrando la necesidad del uso de los datos para la gestión del transporte urbano y la toma de mejores decisiones empresariales en el futuro.

1.4 Estructura

A partir de aquí, el trabajo se estructura de la siguiente manera:

El segundo capítulo, Big Data aplicado a la empresa, explica la evolución que ha tenido el uso de datos masivos o "Big Data", entrando en detalle en los estándares SQL y NoSQL y tecnologías actuales de computación paralela como Hadoop o Spark. También se

explican técnicas de “Machine Learning” de carácter no supervisado. Este capítulo también identifica y describe qué herramientas utilizan las empresas para el almacenamiento, procesamiento y análisis de los datos, intentando expandir los conocimientos del lector sobre estas tecnologías.

El tercer capítulo, Smart Data y el sector de la movilidad urbana, intenta explicar el incremento en el uso y en el almacenamiento de datos, mediante la evolución histórica que han tenido las ciudades con sus problemas de aglomeración de la población, contaminación, tráfico etc. además de estudiar algunas medidas de movilidad que se han tomado en Londres y en Madrid. Este capítulo profundiza en las actividades que realizan las empresas de transporte y movilidad urbana para mejorar sus servicios y la explicación de algunos sistemas que se utilizan tanto en el sector público como en el sector privado.

El cuarto capítulo nos introduce a un caso práctico y real sobre el sistema de bicicletas eléctricas de la ciudad de Madrid, BiciMAD. Se examinan 3,6 millones de viajes efectuados en el año 2018 con carácter exploratorio, se realiza un estudio de asimetría o desbalanceo y mapeado entre estaciones, se estudia la demanda y como se ve afectada por el clima y, por último, se observa la situación de las estaciones en el mes de septiembre de este mismo año, 2018, para conocer su comportamiento a través de este mes y agruparlas en clústeres.

Por último, en el quinto capítulo se extraen las principales conclusiones y se reflexiona y valora el cumplimiento de los objetivos marcados.

2. Big Data aplicado a la empresa

2.1 Evolución histórica del Big Data, bases de datos y el Machine Learning

Antes de adentrarnos en la movilidad en las ciudades o en el caso práctico debemos entender los principales componentes del trabajo, los datos masivos y el aprendizaje automático. Ambos términos podrían ser categorizados bajo el campo de la ciencia de los datos o “Data Science” y la inteligencia artificial.

El término “Big Data” es muy amplio y abarca definiciones similares donde todas tienen algo en común; en este sentido, podríamos referirnos al “Big Data” como el conjunto de datos de tal tamaño y complejidad donde se necesitan programas o aplicaciones informáticas de carácter no tradicional para poder procesarlos adecuadamente en un

período de tiempo razonable. Este campo es por lo tanto una mezcla de ciencia computacional, matemática y, sobre todo, estadística. La popularidad del término se le puede atribuir a John Mashey en 1990 (John Mashey, 1990).

Una definición más moderna, de Charles Fox, en su libro "Data Science for Transport", define esta ciencia como (Fox, 2018):

> *"el campo donde se necesitan herramientas de computación paralela para manejar los datos […] y que esto representa un cambio diferente y definido en la informática utilizada, a través de las teorías de programación paralela, y la pérdida de algunas de las garantías y capacidades realizadas por el modelo relacional de Codd"* (p. 147).

El modelo relacional de Edgar F. Codd, publicado en 1969 cuando trabajaba para IBM, es un modelo para extraer, modelar y relacionar datos, mediante la lógica lingüística de predicados y la teoría de conjuntos, rama de las matemáticas de donde sale, por ejemplo, el famoso diagrama de Venn. El modelo relaciona tuplas, registros o filas con atributos, campos o columnas. Este modelo es el pilar de lo que hoy se conoce como "Relational Database Managament System" (RDBMS) o base de datos relacional (Codd, 1969).

Poco después, en 1974, aparece por primera vez, el lenguaje SQL o "Structured Query Language" que como su nombre sugiere, permite realizar consultas de manera estructurada a una gran base de datos (Chamberlin, 2012). Hoy en día, SQL es todavía el lenguaje de referencia en las empresas, aunque debido al almacenamiento de nuevos tipos de datos no estructurados (imágenes, vídeos o textos etc.), han surgido alternativas denominadas NoSQL o Not-Only-SQL. Curiosamente, este término surgió en 1998 por Carlo Strozzi, aunque fue popularizado por Johan Oskarsson en 2009 cuando organizó una discusión sobre ello en la red social de Twitter. Las bases de datos NoSQL pueden diferenciarse en cuatro grupos; almacenamiento clave-valor, documental, en grafo y orientado a columnas, aunque no entraremos a explicar cada uno de los mismos (Strauch, 2009).

En la actualidad, la mayoría de los softwares más conocidos para el almacenamiento de datos están basados en el modelo de Codd y el estándar SQL, aunque NoSQL ha ganado fuerza en la última década. Los gestores más conocidos que comparten el estándar SQL son Oracle, MySQL, Microsoft SQL Server, PostgreSQL, IBM DB2, Microsoft Access,

SQLite y MariaDB. Los gestores más conocidos que comparten el estándar NoSQL son MongoDB, Apache Ignite, Aerospike, ArangoDB, MarkLogic, Orient DB, CouchDB, Couchbase, Db2, LMDB e InfinityDB. Todas estas organizaciones y empresas ayudan a una mayor eficiencia en el almacenamiento y procesamiento de los datos a todos sus clientes, ya sean empresas de pequeño, mediano o gran tamaño.

Como ya se ha explicado, el propio formato de los datos es importante y estos suelen ser clasificados según su estructura en tres categorías. Datos estructurados, no estructurados y semiestructurados. Los datos estructurados son aquellos que se almacenan en tablas y están muy definidos en formato y número de caracteres. Los datos no estructurados son aquellos que no tienen un formato unificado, los datos son tal y como se recogieron y no se pueden almacenar en tablas. Por último, los datos semiestructurados son aquellos que tienen sus propios metadatos estructurados donde definen objetos y su relación entre ellos (Dedić & Stanier, 2016). En el caso práctico, se utilizarán datos semiestructurados y se convertirán a estructurados para su posterior análisis.

Desde un punto de vista más teórico y de la escalabilidad computacional, el "Big Data" siempre va acompañado de estas características, conocidas como las tres V. Estas fueron inicialmente, volumen, velocidad y variedad, aunque posteriormente se han añadido dos más, veracidad y valor. En 2001, Doug Laney, realizó un informe de investigación para META group, ahora parte de Gartner, donde teorizaba las tres V. En este informe se define volumen como la cantidad de datos generados, velocidad entendida como la rapidez con la que se generan y procesan los datos y la variedad como el tipo de datos utilizados (Laney, 2001). La veracidad y el valor son dos nuevas características que responden a la calidad y la utilidad de los datos con los que se trabajan y que no estaban incluidas en el informe original.

Por otro lado, el aprendizaje automático o "Machine Learning" es otra rama de la ciencia de los datos y la inteligencia artificial muy relacionada con los datos masivos. La continua recolección y el procesamiento de datos ayuda a los modelos de aprendizaje automático a ser cada vez más eficaces, por lo que ambas disciplinas se apoyan la una a la otra. El término aprendizaje automático o "Machine Learning" tiene sus orígenes en 1952, cuando Arthur Samuel crea el primer programa capaz de aprender y jugar a las damas. En la década de 1960 a 1970, numerosos investigadores se interesaron por aplicar la

comunicación humana a la ciencia computacional, lo que dio lugar al procesamiento de lenguajes naturales. En paralelo, otros investigadores pensaron aplicar estos descubrimientos a la biología del cerebro y las neuronas, lo que dio lugar a las redes neuronales artificiales, con el primer modelo denominado "perceptrón" por Frank Rosenblatt en 1960. El perceptrón tenía carencias de orden lógico puesto que no podía representar problemas de carácter no lineal, lo que fue criticado por Minsky and Papert en 1969 (Minsky & Papert, 1969). A lo largo de esta década surgieron otros algoritmos como K-Nearest Neighbors por Cover y Hart en 1967 y K-means, por Lloyd en 1982 y por Forgy en 1965. En 1980, las redes neuronales artificiales resurgieron debido al aumento de la capacidad computacional y, por consiguiente, a poder realizarlas de manera multicapa para tratar problemas no lineales. En estos años, también se desarrollaron los "Expert Systems", que consistían en entrevistar y programar lo descubierto en sistemas con árboles o tablas de decisión. Eran especialmente populares en el sector transporte. Algunos ejemplos son el sistema FRED ("Freeway Realtime Expert System Demonstration") o el sistema FASTBRID ("Fatigue Assesment of Steal Bridges"). En estos años, Judea Pearl crea el concepto de las redes bayesianas. De 1980 a 1990, lo más llamativo fueron los métodos de núcleo o kernel donde destaca especialmente el SVM o Support Vector Machine (Antoniou, Dimitriou, & Pereira, 2018).

En esta sección se ha introducido al lector en dos campos fundamentales para entender el motivo del trabajo. El auge histórico, tanto en el incremento de plataformas de almacenamiento de datos disponibles como en el aumento del uso de algoritmos y modelos para estimar la realidad, deja entrever una tendencia clara. Las empresas y organizaciones que se sumen a estas tendencias tendrán mayor visibilidad a futuro que aquellas que no lo hagan.

2.2 Herramientas de Business Analytics y Big Data en la empresa

En el entorno empresarial, organizaciones y compañías utilizan las siguientes herramientas para el almacenamiento, tratamiento, visualización y análisis de los datos. Este apartado tiene como fin conocer y describir brevemente las tecnologías más utilizadas en la actualidad.

Con el auge de las nuevas tecnologías, la cantidad de datos producidos a analizar es cada vez mayor y de la misma manera, el procesamiento de estos lleva mayor tiempo. Los

sistemas tradicionales RDBMS, la mayoría basados en SQL, no pueden analizar datos semi estructurados o no estructurados como correos electrónicos, mensajes de texto, vídeos o fotografías, por lo que las bases de datos NoSQL se convierten en una gran alternativa. De la misma manera, las empresas encontraron una solución para acelerar el procesamiento de una gran cantidad de datos en 2006. Esta solución es Apache Hadoop, el cual sigue siendo junto con Apache Spark, la tecnología más utilizada. Apache Hadoop es un conjunto de herramientas de código abierto que permite la creación de una red o grupo de ordenadores y la coordinación simultánea de estos para resolver problemas relacionados con el almacenamiento, procesamiento y análisis de datos masivos, lo que se conoce como computación paralela. Estas herramientas son Hadoop Distribution File System (HDFS) para el almacenamiento de datos, MapReduce para el procesamiento de estos y Yet Another Resource Negotiator (YARN) para controlar los ordenadores de la red. Las ventajas principales que ofrece son la velocidad del procesamiento de grandes datos que no se podrían procesar en un solo equipo u ordenador o la escalabilidad de estos sistemas, además de la seguridad que ofrece. Por el contrario, Apache Spark es parecido a Hadoop en funcionalidad, pero tiene una capacidad de procesamiento mayor debido a la utilización de memoria RAM (Hazarika, Ram, & Jain, 2017). Es por esto por lo que las empresas suelen utilizar Hadoop a la hora de almacenar datos mientras que se inclinan por Spark para procesarlos, utilizando una estructura híbrida. Un ejemplo de empresa que ha venido utilizando Hadoop es Facebook, no solamente en el análisis de datos sino en servicios tan importantes para la empresa como la mensajería instantánea (Borthakur, y otros, 2011). En la ilustración 1, se explica el modelo MapReduce. Los datos iniciales (Input Data Files) son partidos en partes y procesados simultáneamente para ser organizados, reducidos y exportados como resultado final (Output Data).

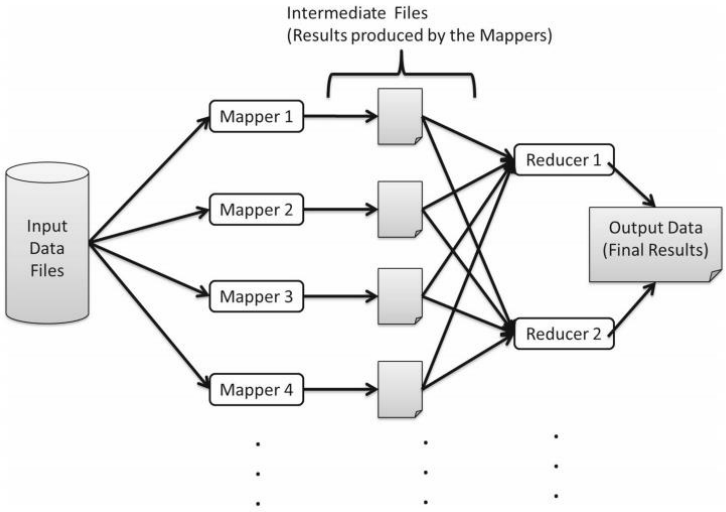


Ilustración 1. Proceso MapReduce para una base de datos utilizando Hadoop

(Stahl, May, Mills, Bramer, & Gaber, 2015).

En el terreno de la visualización de datos, Tableau y PowerBI son las herramientas más utilizadas, aunque también se debería mencionar Qliksense, Microstrategy o Alteryx. Tableau en específico es utilizada por empresas como Coca-Cola, Walmart, Mondelez Internacional, Lenovo o Linked-In. Un caso de mejora en el sector transporte viene de la mano de la aerolínea Southwest, donde utilizando Tableau, es capaz de medir en tiempo real el rendimiento de cómo de puntuales son, calculando la diferencia entre las horas oficiales de llegada y salida planificadas y las reales para cada vuelo. También se recalca que todas estas visualizaciones no son realizadas por el departamento de IT, sino por analistas de negocio dentro de la empresa, lo que sugiere la facilidad de uso que tienen este tipo de herramientas (Tableau Software, 2020). En la ilustración 2, se puede visualizar la interfaz inicial de Tableau con algunos de los gráficos que se pueden realizar.

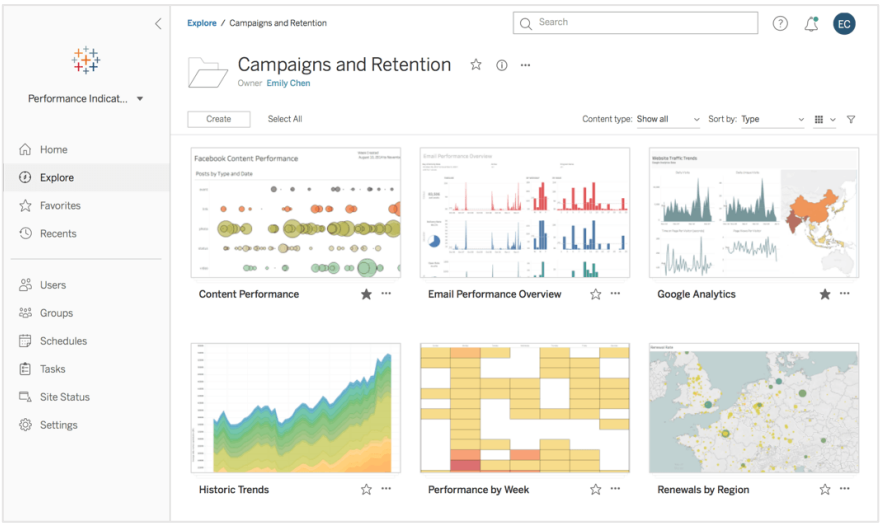


Ilustración 2. Tableau y su entorno *(Tableau Software, 2020)*.

Por último, los lenguajes de programación para el análisis de datos más utilizados, según el ranking IEEE Spectrum, son Python y R (Cass, 2019). Python es más versátil y no solo se aplica al terreno de la ciencia de los datos, por lo que es el lenguaje más popular en la actualidad. Por el contrario, R es un lenguaje desarrollado por estadísticos y se centra en el análisis y creación de modelos estadísticos. Con el tiempo, se han desarrollado interfaces que permiten la mejor gestión de estos, siendo Jupyter Notebook y RStudio las más utilizadas. Ambos lenguajes son utilizados por grandes empresas; por ejemplo, en la página web de RStudio, se puede ver que empresas de todo tipo, como General Electric, Honda, Accenture, Samsung, Banco Santander, NBC, Ebay o Waze, utilizan R en el día a día (RStudio, 2020). De la misma manera, en la fundación Python hay un apartado de éxitos en empresas donde dejan entrever cómo se utiliza en Philips o Honeywell (Python.org, 2020).

2.3 Machine learning: aprendizaje no supervisado

Algunos de los principales algoritmos del aprendizaje automático se han nombrado en el apartado 2.1. Debido al gran número de algoritmos, en este apartado solo se explican los algoritmos utilizados en la parte práctica. No obstante, se deben diferenciar antes las dos clases de aprendizaje automático más conocidas: el aprendizaje supervisado y el no-supervisado.

El conjunto de datos en el aprendizaje supervisado son un conjunto de datos con etiqueta. El objetivo de un algoritmo de aprendizaje supervisado es utilizar un conjunto de datos para producir un modelo que tome un vector de características X como información de entrada y que pueda producir una salida que permita deducir la etiqueta de este vector. Por ejemplo, el impago de un deudor en un crédito o salida vendrá determinado por las características financieras y personales de este deudor o entrada. Por otro lado, el aprendizaje no supervisado se caracteriza por tener un conjunto de datos sin etiquetar que utiliza el mismo vector de características X y realiza una transformación sobre este, resultando en otro vector o valor con el fin de resolver un determinado problema. Este aprendizaje carece de una variable dependiente dada a priori. Un ejemplo de algoritmo de aprendizaje no supervisado es el "clustering".

El "clustering" se refiere a técnicas de agrupación para encontrar subgrupos o clústeres en un conjunto de datos, de tal forma que las observaciones dentro de un mismo grupo sean bastante similares entre sí, y las observaciones de diferentes grupos bastante diferentes entre sí (James, Witten, Hastie, & Tibshirani, 2013). En "clustering" y siguiendo con el ejemplo financiero, se podría determinar el tipo de cliente o clúster al que un determinado cliente en el conjunto de datos pertenece en comparación al resto, agrupándolos por la similitud de sus hábitos financieros y de consumo (Burkov, 2019). En realidad, hay otros dos tipos de aprendizaje automático: el aprendizaje semi supervisado y el aprendizaje de refuerzo, aunque no son explicados en detalle, ya que quedan fuera del alcance de este trabajo.

A continuación, se explican los algoritmos utilizados en la parte práctica, todos ellos de carácter no supervisado y, más concretamente, algoritmos de "clustering". En el anexo I encontrarán otras aplicaciones en las que se pensaron inicialmente, pero debido a la complejidad requerida y a la falta de datos, se ha optado por no incluirlas o explicarlas en este apartado, con el fin de no confundir al lector.

2.3.1 Clustering por K-medias/K-means

El algoritmo K-medias es el algoritmo más utilizado dentro de la familia de "clustering". Este algoritmo pertenece a la familia de formación de clústeres por partición en k grupos. Otros algoritmos en este grupo son K-medoids o PAM y CLARA o "Clustering Large Applications", aunque nos centraremos en explicar el primero y más conocido de ellos.

En K-means, k o el número de grupos o clústeres es especificado de antemano, mediante una serie de indicadores estadísticos (alrededor de 30) que nos indican el número óptimo de grupos o clústeres para el conjunto de datos dado. K-means clasifica observaciones en k grupos, donde las observaciones dentro del mismo clúster son muy similares entre sí, y donde las observaciones de diferentes grupos son tan disímiles como fuera posible. El nombre K-means viene dado debido a que cada grupo está representado por su centroide, que corresponde a la media de puntos asignados al grupo (Kassambara, 2017).

En la ilustración 3 se puede observar el proceso y los centroides finales. De una manera práctica, el algoritmo comienza seleccionando k observaciones y calcula los centroides iniciales de los grupos o clústeres. Cada una de las observaciones se asigna al centroide más cercano, medido a través de la distancia euclidiana/euclídea generalmente. Después de este paso de asignación a los grupos o clústeres, el algoritmo calcula de nuevo el centroide o valor medio del clúster (Kassambara, 2017). Este proceso se repite cuantas veces se pueda, hasta el momento en el que una iteración o repetición de más no modifique los centroides o valores medios de los clústeres.

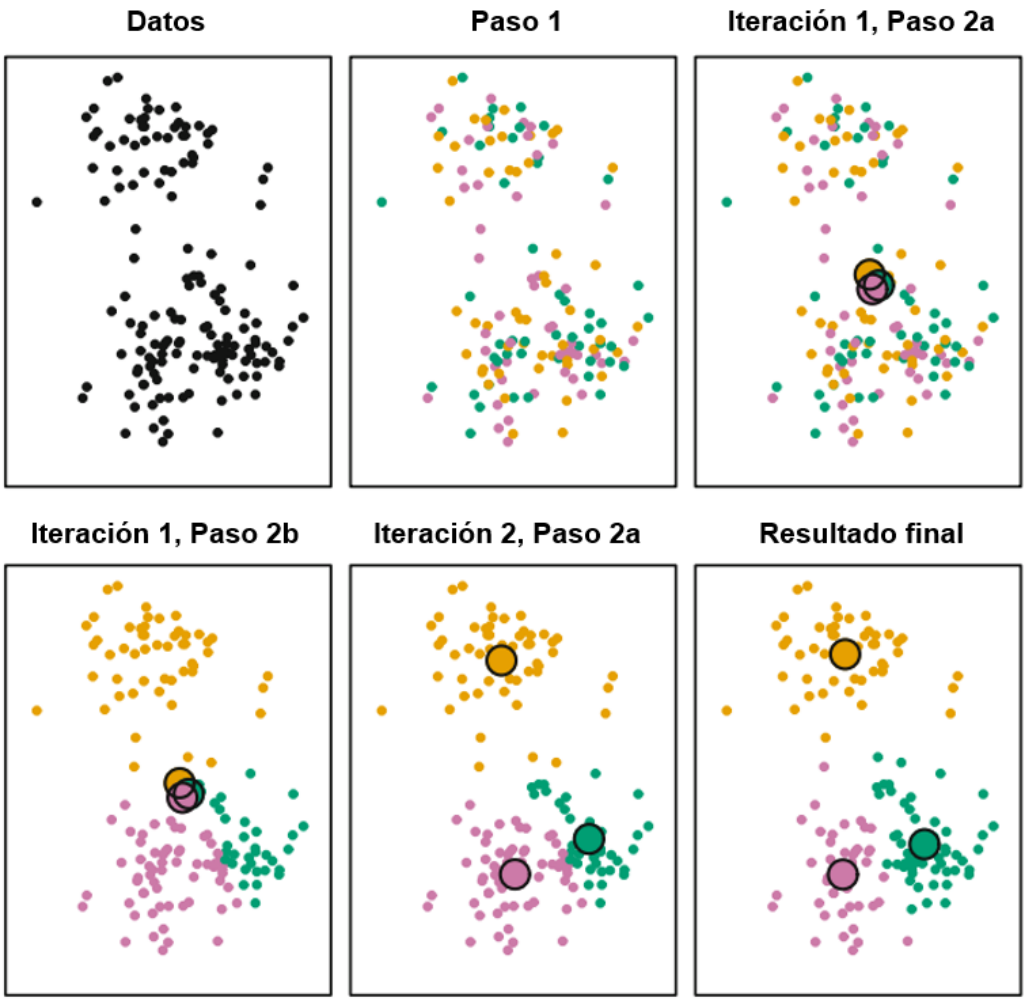


Ilustración 3. Proceso de agrupación por K-means y centroides

(James, Witten, Hastie, & Tibshirani, 2013).

La expresión del algoritmo K-means se define como;

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

donde Xi es una observación en los datos que pertenece al grupo o subconjunto Ck y μk es el valor medio de los puntos asignados al clúster Ck. Cada observación es asignada a un clúster de manera que la suma de los cuadrados de la distancia de la observación al centroide sea la mínima posible. La distancia más utilizada para realizar estos cálculos es la distancia euclídea. La variación total dentro de cada clúster es definida como;

$$\sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

La suma de cuadrados total dentro del clúster mide la compactación de este, por lo que queremos que sea lo más pequeña posible.

2.3.2 Clustering jerárquico

El "clustering" jerárquico es otra forma de realizar grupos o clústeres. El algoritmo está basado en formar grupos basados en cómo de similares son las observaciones, por lo que no se predefine un número de clústeres. El "clustering" jerárquico se puede dividir en dos tipos, aglomerativo y divisivo. El tipo aglomerativo, al inicio, clasifica cada observación como un propio clúster (hoja o leaf) y progresivamente, los clústeres más similares se van uniendo hasta formar un gran clúster (raíz o root). El tipo divisivo es justo lo contrario, donde se forma un gran clúster y se va dividiendo, basado en como de diferentes sean los clústeres entre sí. Este proceso se puede ver en un gráfico denominado dendograma (Kassambara, 2017).

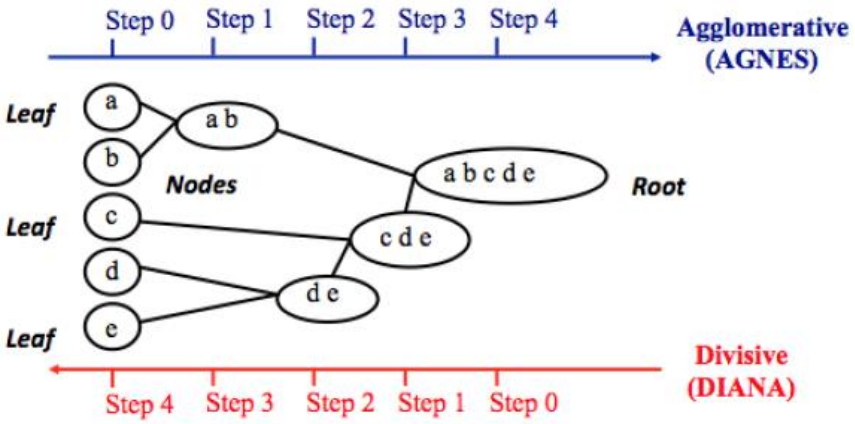


*Ilustración 4.*Tipos de "clustering" jerárquico que existen, su metodología y dendograma horizontal (Kassambara, 2017).

En "clustering" jerárquico, la distancia más común para calcular como de similares o dispares son los puntos u observaciones es la distancia euclídea, distancia ordinaria desde un punto A a un punto B. No obstante, también es común utilizar el método de Ward donde el criterio fundamental es minimizar la variación total dentro del grupo o clúster.

Además, es posible realizar "clustering" de series temporales, con el objetivo de agrupar las observaciones por la similitud o diferencia en la forma (que presentan sus mediciones en el tiempo) en grupos o clústeres. Por lo que en estos casos se utilizan medidas de distancia más específicas, trabajando generalmente con la distancia DTW o Dynamic Time Warping. Esta idea fue introducida y descrita por Bernt y Clifford en el año 1994, con resultados prometedores en su informe "Using Dynamic Time Warping to Find Patterns in Time Series" (Berndt & Clifford, 1994). Otros estudios han investigado sobre las fortalezas y limitaciones de la distancia DTW y es sabido que, al comparar series temporales, esta medida presenta una efectividad superior a la euclídea debido a que tiene en cuenta distorsiones temporales en la variable temporal (ya sean mediciones por hora, fecha, día etc.) (Ratanamahatana & Keogh, 2004); razón por la cual la utilizaremos en el caso práctico cuando realicemos "clustering" por series temporales. La efectividad de utilizar la distancia euclídea frente a DTW al analizar una serie temporal se puede observar en la ilustración 5.

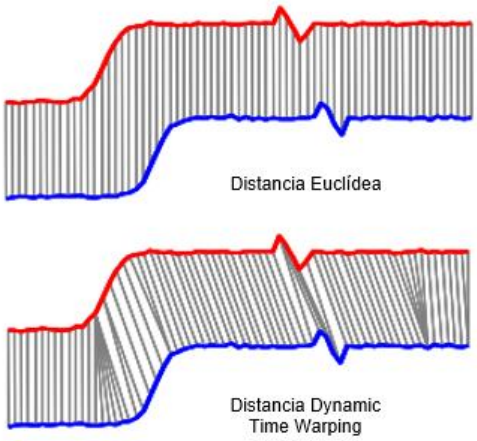


Ilustración 5. Efectividad de la distancia Euclídea frente a la distancia DTW en la comparación de dos series temporales (azul y roja) (Ratanamahatana & Keogh, 2004).

Finalmente, a la hora de utilizar "clustering" jerárquico, se han de tener en cuenta los criterios de enlace más conocidos:

1. Complete: busca la máxima disimilitud "inter-cluster" (entre los dos o más clústeres). Calcula todas las diferencias por pares entre las observaciones del clúster A y el clúster B. Registra la mayor de estas diferencias.

2. Single: busca la mínima disimilitud “inter-cluster” (entre los dos o más clústeres). Calcula todas las diferencias por pares entre las observaciones del clúster A y el clúster B. Registra la menor de estas diferencias.
3. Average: busca la disimilitud media “inter-cluster” (entre los dos o más clústeres). Calcula todas las diferencias por pares entre las observaciones del clúster A y el clúster B. Registra la media de estas diferencias.
4. Centroid: Busca la diferencia entre el centroide para el grupo A (una media vector de longitud p) y el centroide del grupo B.

(James, Witten, Hastie, & Tibshirani, 2013). Al realizar “clustering” en el caso práctico, se ha utilizado el método “average” con el fin de tener en cuenta todas las observaciones/series temporales en el dendograma resultante.

2.3.3 Métodos para la identificación de Clústeres Óptimos

Existen una treintena estadísticos de métodos para identificar el número de clústeres o grupos a formar para un conjunto de datos. Sin embargo, en esta sección se explican únicamente el método de Elbow y Silhouette, al ser considerados los más utilizados en la práctica. Ambos estiman de diferente manera el número de clústeres óptimos para un conjunto de datos determinado.

1. El método Elbow está basado en los principios del “clustering” por partición y calcula la varianza total “intra-cluster” (entre las observaciones de un mismo clúster) en función del número de clústeres. Esto es, intenta encontrar el número de clústeres con el que la varianza dentro de cada clúster o la suma de los cuadrados dentro de cada clúster sea minimizada (Kassambara, 2017). El número de clústeres es elegido basado en el momento en el que añadir otro clúster no supone una reducción en la varianza total tan significativa.
2. El método Silhouette es una medida que estima cuán similar es un elemento a su propio grupo (cohesión) en comparación con otros grupos (separación). El gráfico resultante muestra los valores promedio de la silueta de los elementos y ayuda a determinar el número "óptimo" de grupos. El objetivo es maximizar la media del índice silueta. El índice silueta mide como de buena o mala es la asignación de una observación a un clúster y toma valores de -1 a 1, siendo los valores cercanos

a 1 indicativo de que la observación se ha asignado correctamente. (Kassambara, 2017).

En la ilustración 6, se formarían 4 clústeres utilizando el método Elbow y únicamente 2 utilizando el método Silhouette.

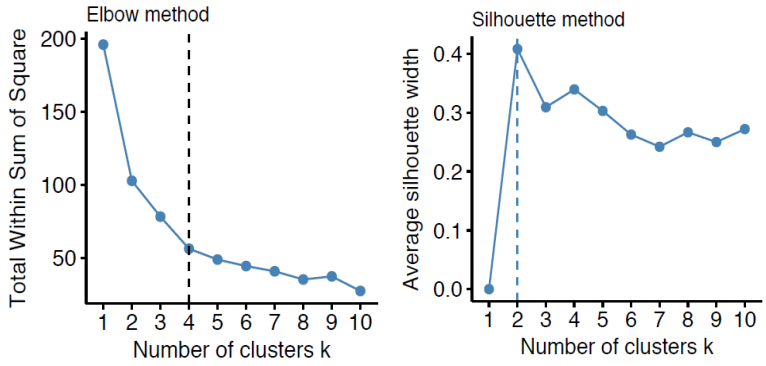


Ilustración 6. Método de Elbow y Silhouette *(Kassambara, 2017)*.

3. Smart data y el sector de la movilidad urbana

3.1 Principales problemas de la movilidad urbana y las ciudades

Para comprender por qué se buscan soluciones a la movilidad urbana mediante el uso de datos en las grandes ciudades, debemos realizar un recorrido histórico a los principales hechos demográficos en las últimas décadas.

La congestión y aglomeración de la población en las ciudades es un fenómeno sin precedentes que se aceleró con la primera revolución industrial y que se acentuó aún más en la segunda mitad del siglo XX de manera global (Achával, 1950). Es el fenómeno denominado "Éxodo rural", donde la población del campo se trasladaba a las grandes ciudades en busca de mejores oportunidades laborales. El Instituto Nacional de Estadística de España (INE) junto con el diario El Confidencial van incluso más allá, y describen los nuevos flujos migratorios en España y en el mundo como éxodos urbanos, de ciudades de pequeño o mediano tamaño a aquellas de gran tamaño. En este análisis desde el año 1988 hasta 2018, se pueden observar varias tendencias, como el incremento de las migraciones de capitales de ciudades como Almería, Pamplona, Santander, Sevilla, Valencia o Zaragoza a las grandes ciudades de Madrid y Barcelona o el significativo aumento de la población en Madrid. No obstante, el dato más interesante es la

concentración de los empleos con alto nivel tecnológico, donde Madrid acumula el 40% de todos los empleos de esta categoría creados en la última década. Esto obliga a los jóvenes y estudiantes a migrar a la Comunidad de Madrid o a Barcelona, ya que es donde residen estas oportunidades laborales (Jorrín, y otros, 2019).

A nivel más general, la globalización ha permitido un aumento generalizado de la población en las ciudades. El INE también desvela los datos del crecimiento de población total en España desde 1971 y el crecimiento de la población de Madrid desde 1996. En Madrid, en el año 1996, había 5.022.289 personas y en 2019, 6.663.394, lo que supone un aumento porcentual del 32,68% (INE España, 2019). En España, en el año 1996, había 39.808.374 personas y en 2019, 47.100.396, lo que supone un aumento porcentual del tan solo 18,32% (INE España, 2019). Estos números dan a entender que ha habido una concentración de la población en las urbes urbanas, pero también pueden advertir que, en el futuro, la población crecerá y se aglomerará, todavía más, en estas grandes ciudades. Estas congestiones en las ciudades dan cabida a nuevos retos en la movilidad urbana, desde lidiar con el tráfico de las carreteras, construir una red de transporte público más eficiente o disminuir la concentración de gases de efecto invernadero a través de medidas políticas y/o económicas.

Probablemente, el ejemplo más conocido de limitación a la circulación sea la tasa de congestión o "congestión charge" impuesta por Transport for London en febrero de 2003, órgano gubernamental de la ciudad de Londres en transporte y movilidad. Esta medida cobra a todo vehículo que circule dentro de una serie de zonas centrales de Londres un importe de 11,50 libras por día, aunque este importe disminuye para los residentes de estas zonas (Transport for London, 2003). Los ingresos de esta medida económica van destinados a la mejora del transporte público de Londres y contribuyen a la reducción del tráfico y al impulso del uso del transporte público en la ciudad. Los vehículos híbridos y eléctricos están actualmente exentos del pago de esta tasa, pero a partir de octubre de 2021, estarán únicamente exentos los vehículos eléctricos (Transport for London, 2019). Esto es debido a que el 8 de abril de 2019, el alcalde de Londres lanzó la iniciativa ULEZ o "Ultra Low Emission Zone", una serie de restricciones económicas a aquellos vehículos que se adentrasen en ciertas zonas de la ciudad y que no cumpliesen con los estándares de etiqueta de emisiones (Euro 3, Euro 4, Euro 6 etc.), con el fin de mejorar la calidad del aire y, por consiguiente, la salud de todos los londinenses (Mayor of London, 2019).

Otras iniciativas de Transport for London para mejorar la movilidad sostenible de la ciudad son las conocidas Santander Cycles, la red de bicicletas eléctricas de Londres con más de once mil bicicletas y ochocientas estaciones a julio de 2018 (Transport for London, 2018). El proyecto inicial de estas bicicletas se determinó en 140 millones de libras de inversión a lo largo de seis años y se estimó que podría ser una fuente de ingresos a largo plazo, cubriendo el coste operativo en únicamente tres años. Esto era una noticia muy positiva, teniendo en cuenta que Transport for London sigue y seguía perdiendo dinero por cada viaje en metro o autobús (Whitehead, 2010). Por último, desde la iniciativa privada, la polución en el aire de Londres ha incentivado la transición al vehículo híbrido de empresas como GLH en 2005, una empresa de taxis híbridos y eléctricos que ha crecido hasta convertirse en un referente del transporte sostenible en Londres. En 2018, GLH adquirió eConnect Cars, una empresa de taxis que también utiliza vehículos eléctricos (GLH, 2018).

Desde una perspectiva más cercana, en Madrid, que goza de mejor calidad de aire que Londres, también se han impulsado medidas más severas de carácter medioambiental y de movilidad. La aprobación de la zona denominada Madrid Central en 2018 es el más claro ejemplo de estas medidas. Madrid Central es el área de bajas emisiones de la ciudad de Madrid, con un amplio perímetro de 472 hectáreas. Restringe el acceso de todo tipo de vehículos diésel o gasolina salvo para los residentes, sin distinción alguna por la etiqueta europea de emisiones o por la etiqueta medioambiental de la DGT y exime a todo vehículo eléctrico o híbrido (Ayuntamiento de Madrid, 2018). El 30 de septiembre de 2019, el ayuntamiento anunciaba un plan alternativo a esta zona, con nombre Madrid 360, que permitirá el acceso a vehículos con el distintivo medioambiental C o superior siempre y cuando estén ocupados por dos o más personas (Belver & R. Roces, 2019).

Otras medidas que impulsan las ciudades españolas son las que rodean a las distintas zonas de aparcamiento en la calle, siendo estas categorizadas en los colores verde y azul. Las zonas verdes cuestan menos dinero para los residentes, con el fin del uso del transporte público entre el centro y la periferia de Madrid. Asimismo, estas zonas verdes establecen un límite de tiempo para aquellos que no sean residente de la zona en la que se ha aparcado, fijado en dos horas aproximadamente. Una vez más, en Madrid, el precio a pagar por aparcar en ambas zonas parte de una tarifa base, donde se aplican reducciones o recargos en función de la tecnología del vehículo y la etiqueta medioambiental de la

DGT. Las motocicletas, ciclomotores, bicicletas y vehículos impulsados por motor eléctrico pueden aparcar de manera gratuita (RACE, 2019).

La efectividad de estas medidas de carácter económico, político o social no se puede medir sin el uso de los datos. La tecnología y en específico el "Big Data" no solo nos ayuda a realizar acciones de una forma más rápida sino también a conocer si estas acciones han sido efectivas con el paso del tiempo. Por ejemplo, tras la anteriormente mencionada implementación de la ULEZ en Londres, el ayuntamiento ha sido capaz de saber cuán efectivas han sido estas medidas mediante la recolección y el análisis de datos antes y después de implementar la medida. Desde la implementación de la medida hasta el 21 de octubre de 2019, la ULEZ ha disminuido 13.500 vehículos en los barrios más céntricos de Londres, reduciendo así el tráfico. No parece ser una gran cifra, hasta que se miran los aumentos y disminuciones de contaminación en el aire, que se han conseguido reducir en un 1/3 del total. Esto es debido a que los vehículos más contaminantes han sido los más perjudicados donde el porcentaje de vehículos que cumplen con los estándares de emisiones de la ULEZ ha aumentado de un 61% del total en marzo de 2019 al 77% en octubre de 2019 (Mayor of London, 2019).

Esta sección del documento ha cubierto una descripción de los principales flujos migratorios hacia las ciudades, además de estudiar los principales problemas de la movilidad urbana a través de la sostenibilidad. Desde el punto de vista del analista de datos, el mayor número de población concentrada o mayor densidad de población en las ciudades da lugar a un mayor número de observaciones para cada estudio. Esto lleva a cantidades de datos masivos por analizar y a la creación de herramientas para dicho cometido. La iniciativa privada y pública busca entonces mejorar la movilidad mediante el uso y análisis de los datos recolectados, surgiendo nuevos modelos de negocio como los de economía colaborativa o iniciativas públicas para la mejora del transporte público.

3.2 El uso de los datos en el sector de la movilidad urbana

Una vez descritas algunas de las medidas tomadas y de las ventajas que se pueden obtener con el uso de los datos, debemos adentrarnos en el sector de la movilidad. Para cumplir con este cometido, se introduce al lector a un cuadro de inversión actual sobre las tecnologías más populares en la movilidad urbana en la ilustración 7. Del mismo modo,

se reflexiona sobre los principales problemas en las ciudades (gestión de tráfico, sostenibilidad, vehículos autónomos, etc.) y cómo las tecnologías disponibles intentan mitigarlos e incluso solucionarlos. Por último, se introducen las soluciones Mobility-as-a-Service.

De nuevo, la ilustración 7, traducida al castellano, muestra cómo la tecnología que ha crecido más respecto a 2010 es “E-hailing”. Este término es definido como la capacidad de pedir un vehículo a través de medios digitales, como un teléfono inteligente o un ordenador. Por otro lado, la única tecnología donde la inversión decrece es el reconocimiento de voz, ya que se han dado grandes avances en los últimos años. Lo más significativo del gráfico es la gran inversión en el vehículo autónomo (Möller, Padhi, Pinner, & Tschiesner, 2019).

Las empresas pertenecientes al grupo “E-hailing” más conocidas son Uber, Lift, OLA, FreeNow y Cabify, que utilizan datos para entender las necesidades y expectativas de sus clientes, según un informe de Arthur D. Little. Este último informe explica las oportunidades que surgen en el sector mediante el uso de estas plataformas como la distribución óptima de recursos, la reducción de costes y la optimización de precios mediante algoritmos y recalca la importancia del sector público en las ciudades y el papel regulatorio que tiene el estado para controlar, por ejemplo, las aglomeraciones en el tráfico o la seguridad de los ciudadanos, además de asegurar que el taxi autónomo es el futuro de la movilidad (Audenhove, y otros, 2020).

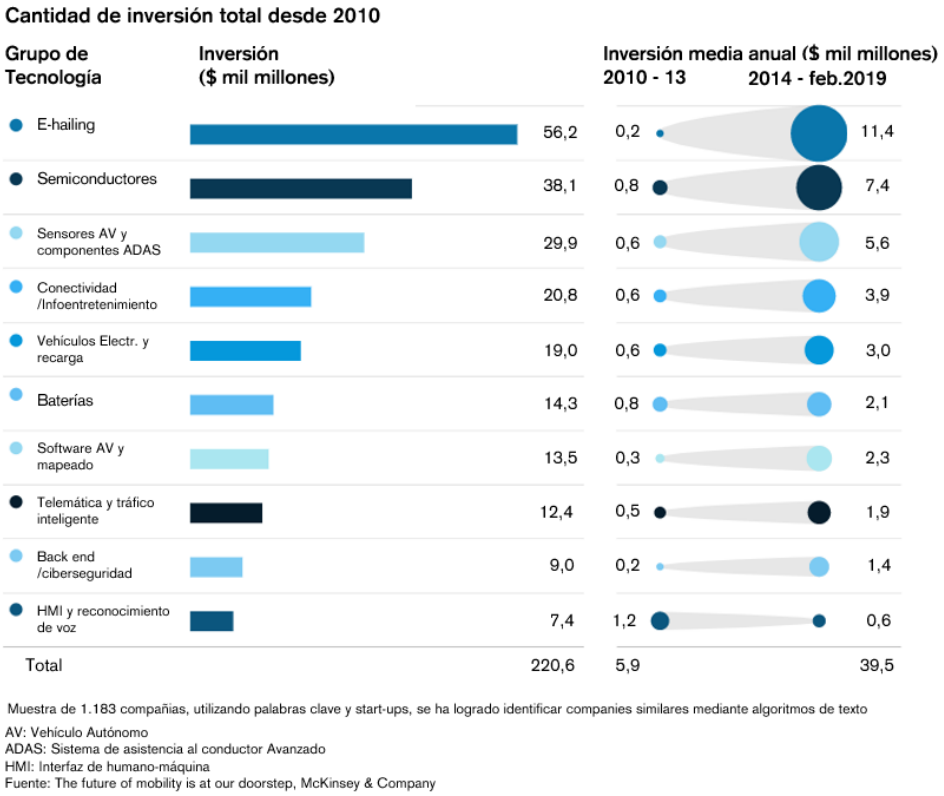


Ilustración 7. Inversión estimada en el sector movilidad *(Möller, Padhi, Pinner, & Tschiesner, 2019)*.

A nivel europeo, la iniciativa "Transforming Transport" financiada por la Comisión Europea con un importe de 18,7 millones de euros y coordinada por la empresa española de sistemas, Indra, pretende implementar nuevos algoritmos basados en "Big Data" para mejorar la movilidad y particularmente, su gestión y servicio. La iniciativa une así a los principales operadores de transporte europeos en proyectos de muy diversa índole a través de toda Europa (Indra Sistemas, 2017). Algunos de estos proyectos se centran en la mejora de temas recurrentes en la movilidad como la sostenibilidad, la seguridad vial o el desarrollo de vehículos autónomos.

3.2.1 Gestión de tráfico y Sistemas Inteligentes de Transporte (ITS).

El tráfico es uno de los mayores problemas de las grandes ciudades. Dos investigadores de la universidad de Chalmers realizaron un estudio, mediante datos obtenidos por la API de HERE Traffic (una de las plataformas líderes en movilidad con mediciones y soluciones basadas en el uso de datos), del tráfico de 45 ciudades, además de formar clústeres o grupos con las ciudades según el comportamiento de series temporales de tráfico de 24 horas, parecido a lo que se realizará con el caso práctico, presentado en el apartado 4 de este trabajo. También accedieron a varios sensores de tráfico ITS

("Intelligent Transport System") de la Agencia Sueca de Transporte con el fin de comparar los datos de la API de HERE Traffic, basados en datos de sensores en vehículos, teléfonos inteligentes, dispositivos de navegación personal, sensores de carretera y coches conectados e informes de incidentes y accidentes públicos, con estos sensores reales para unas determinadas vías. El resultado fue una gran similitud entre ambas mediciones, lo que sugiere una alta fiabilidad por parte de estas plataformas en la medición del tráfico (Verendel & Yeh, 2019).

Otro estudio discute las aplicaciones del "Big Data" a las ciudades, donde se explica la posibilidad de implementar un sistema inteligente de semáforos interconectados, para aligerar la congestión del tráfico y dar más información sobre la velocidad promedio de los vehículos en la vía, la densidad del tráfico, tiempo de espera medio en el semáforo y posibles aglomeraciones. El sistema realiza decisiones basado en estos parámetros y manda instrucciones a otros semáforos donde cuanta más información recibe, mejor decide. Esta iniciativa se realizó en la ciudad de Pittsburgh, Pennsylvania junto a Traffic21, instituto sobre el estudio del transporte de la Carnegie Mellon University, donde se redujeron las emisiones de gases un 20% y se aligeró la congestión de la ciudad (Nuaimi, Neyadi, Mohamed, & Al-Jaroodi, 2015).

3.2.2 Sostenibilidad: Eficiencia energética a través del dato

Respecto a la sostenibilidad, se han comentado en anteriores apartados las iniciativas que se han tomado en Londres o Madrid. En este apartado expandimos esta idea, con otras aplicaciones que se han dado o podrían darse en medios de transporte específicos. En este apartado sobresale la implementación de Accenture en el circuito de ventilación del metro de Madrid que ha ahorrado 1.800 toneladas de emisiones CO2 y ha supuesto una reducción del 25% en los costes energéticos en 2019 (Sánchez & Díaz, 2019):

> *"El sistema utiliza un algoritmo de optimización capaz de movilizar grandes cantidades de datos para obtener todas las combinaciones posibles en cuanto a temperatura del aire, arquitectura de la estación, frecuencia de trenes, carga de pasajeros y precio de la electricidad a lo largo del día. El algoritmo utiliza datos tanto históricos como simulados, y tiene en cuenta la temperatura externa y la del subsuelo durante las siguientes 72 horas. Además, como utiliza "Machine*

Learning", el sistema va mejorando a la hora de predecir el balance óptimo para cada estación de la red a lo largo del tiempo" (p.1).

Por otro lado, Accenture junto con Telia Company, analizó patrones de movimiento a través de teléfonos inteligentes de los usuarios de una compañía de autobuses al sur de Estocolmo. Los resultados mostraron que los viajeros gastaban demasiado tiempo en una ruta porque había demasiadas paradas y la compañía de autobuses añadió un autobús más, para aminorar la cantidad de viajeros. De la misma manera, los desplazamientos matutinos disminuyeron un 16% los viernes, por lo que la compañía de autobuses ajustó el número y la frecuencia para optimizar el uso en estos días. La plataforma utilizada para realizar estos análisis ha sido utilizada recientemente para determinar la construcción de una línea de metro entre Espoo y Helsinki, donde se ha conseguido realizar un replanteamiento de las conexiones entre las estaciones según los patrones de movimiento de los usuarios. Esta línea de metro también ha reducido la contaminación en 13 toneladas de CO2 al día, debido a la disminución de un 8% del tráfico entre ambas ciudades (Accenture, 2019).

En la misma línea, otro estudio discute la implementación de las tarjetas de trasporte público con todas las ventajas y limitaciones que suponen, permitiendo mediante el análisis de datos, identificar los flujos de personas entre estaciones. Este estudio también propone para ciudades como Beijing, una restricción de energía por habitante, mediante un algoritmo que calcula los costes de energía por cada viaje dentro de un área, determinados por el tipo de vehículo en el que se viaja y las condiciones de tráfico en caso de elegir un vehículo privado, pero asignando a los vehículos no-motorizados el valor cero. Cada individuo puede mirar su cuota de energía restante para variar el método de viaje y mantenerse dentro de la cuota (Wang & Moriarty, 2018). Este método, aunque efectivo para ciudades con un índice de polución muy alta, no me satisface y sería una regulación muy estricta y/o privativa.

3.2.3 Vehículos autónomos y seguridad en la vía

El vehículo autónomo eléctrico convencional utiliza sensores por los que capta miles de datos y con los que decide, a través de redes neuronales profundas, las decisiones para llegar del punto A al punto B de la manera más eficiente posible, previene accidentes e intenta dar solución a una movilidad más cómoda, eficiente y limpia en emisiones.

Algunas de las empresas más avanzadas en el desarrollo del vehículo autónomo son Waymo (subsidiaria de Google), GM Cruise (subsidiaria de General Motors), ArgoAI (subsidiaria de Ford), Tesla, Baidu, Drive.ai (subsidiaria de Apple), BMW, Daimler, Volkswagen Group, Uber y Lift.

En específico, la compañía Waymo capta y recrea su entorno mediante un sensor LIDAR o Light Detection and Ranging, además de la utilización de cámaras y radares. Este sistema ilumina el objetivo con un láser y mide el reflejo de la luz con un sensor para determinar la distancia a los objetos e identificarlos en todo momento. Los sensores LIDAR son un referente y, de hecho, pueden ver a mayor distancia que sensores convencionales, lo que previene accidentes, pero no todos, como es el caso del accidente mortal de Uber en 2017 (Templeton, 2019). La mayoría de estas empresas utilizan redes neuronales no supervisadas como método para la toma de decisiones, enseñando al sistema con horas de conducción real (Tian, Pei, Jana, & Ray, 2018). En la ilustración 8, se puede observar, en una zona de obras, lo que ve en coche autónomo en la izquierda, frente a la visibilidad real a la derecha.

Ilustración 8. A la derecha, lo que se ve en la calzada; a la izquierda, el vehículo de Waymo y lo que está viendo en una zona de obras *(Waymo, 2018)*.

Este último debate nos abre la puerta a la pregunta ¿Qué tan seguros son estos vehículos? Por ley, toda empresa que prueba vehículos autónomos en el estado de California debe dar una serie de datos al gobierno basado en su número de desconexiones (cuando el vehículo falla y desconecta su sistema de conducción autónoma ante condiciones

adversas) por millas recorridas. Un estudio publicado por el Centro Nacional de Información Biotecnológica analiza esta información desde septiembre de 2014 hasta enero de 2017, determinando una tasa de desconexión y accidentes extremadamente baja. La mayoría de los accidentes fueron causados por otras personas por un impacto en la parte trasera del vehículo autónomo en intersecciones y a velocidades muy bajas. Finalmente, se estima que un coche autónomo capaz de adaptarse a cualquier situación podría alcanzarse alrededor del año 2030 (Favarò, Nader, Eurich, Tripp, & Varadaraju, 2017).

Para alcanzar esta meta, no solo deben mejorar los vehículos sino también, las infraestructuras. Volviendo al ejemplo de la iniciativa "Transforming Transport", uno de los proyectos piloto implementados en España consiste en una autopista inteligente denominada AUSOL Load Balancing Pilot. Este proyecto se lanza en la AP-7 donde se implementarán soluciones para entender los factores del tráfico en un tramo de 96 kilómetros, con el fin de manejar y entender los flujos de tráfico, optimizar las operaciones de mantenimiento y aumentar la seguridad, reduciendo el número de accidentes producidos. Todo esto, estará basado en 20 fuentes de datos, tanto de carácter de tiempo real como de carácter histórico (Transforming Transport, 2018).

Finalmente y con la crisis del COVID-19, Imotion Analytics va a implementar el uso de modelos artificiales en las cámaras de varias estaciones de Renfe, capaces de reconocer mediante redes neuronales, las características demográficas de las personas y conocer los principales flujos en horas de alta demanda, con el fin de imponer un control más estricto y hacer frente a la pandemia evitando aglomeraciones (Renfe, 2020).

3.3 Proveedores de soluciones Mobility-as-a-Service (MaaS)

Los gobiernos, al ver la población crecer en las ciudades y con ello los problemas, optan por recopilar datos y analizarlos de manera conjunta, con el fin de operar de manera más eficiente. Esto da cabida a las soluciones Mobility-as-a-Service o movilidad como servicio. Por ejemplo, en 2014, el ayuntamiento de Río de Janeiro, debido a las quejas recibidas por la falta de seguridad y las congestiones de tráfico, colaboró con IBM para combinar todos los medidores y datos de treinta instituciones públicas de la ciudad. Estos datos son posteriormente procesados por un algoritmo obteniendo recuentos y gráficos en

tiempo real. Esta iniciativa se denominó “Centro De Operacoes Prefeitura Do Rio” o Centro de Operaciones del Ayuntamiento de Río y sigue operando en la actualidad. (Kitchin, 2014).

Con el uso de los datos en la industria del transporte y sus ventajas, nace la necesidad de simular flujos de tráfico en las grandes ciudades e incluso medirlos en tiempo real con soluciones de bajo coste. Las herramientas más conocidas para realizar simulaciones de tráfico son Eclipse Simulation of Urban MObility (Eclipse SUMO, 2020) y PTV Vissim (PTV Group, 2020), que ayudan a probar la eficiencia de soluciones de tráfico antes de ser implementadas. Por otro lado, el futuro de la movilidad recae sobre las plataformas de medición en tiempo real denominadas empresas de “Data Analytics” en movilidad y/o proveedores de soluciones MaaS (Mobility-as-a-Service). Entre ellas, destacan empresas internacionales como Moovit (subsidiaria de Intel), Citymapper y HERE Technologies (propiedad de Audi, BMW y Daimler). Estas aplicaciones tienen otras funcionalidades; como agregar todos los medios de transporte de tal forma que los usuarios puedan pedirlos o reservarlos, pagos de bonos y tiques de distintos medios de transporte, recolección y análisis de datos de flujos de transporte y personas etc. (Moovit, 2020).

Una pregunta frecuente es ¿Cómo estas plataformas internacionales captan datos en tiempo real de varias ciudades y cómo pueden estimar el tiempo restante para que un autobús o metro llegué a la parada o estación? Una empleada en Moovit define datos en tiempo real como una predicción inteligente basada en datos de los tiempos de llegada a distintos puntos de una ruta. De forma más específica, Moovit tiene dos maneras de conseguir realizar esta predicción; la primera es integrar los tiempos estimados de llegada que son calculados y compartidos por la agencia de transporte o movilidad de la ciudad observada. La segunda forma se da cuando la agencia de transporte o movilidad de la ciudad observada da la posición geográfica del vehículo en GPS y Moovit calcula los tiempos estimados de llegada para las demás estaciones utilizando su algoritmo especializado. Cuando un usuario abre la aplicación Moovit, el usuario manda una petición al servidor, que es recibida en forma de datos estáticos con sus datos, la parada o estación buscada y eventos a considerar en la ruta. La información recibida por el usuario es actualizada hasta que este cierre la aplicación. Cuando la ciudad no cuenta con una infraestructura para dar este tipo de datos en tiempo real, el producto Moovit TimePro es una buena alternativa. Se trata de una solución en la nube para todos los autobuses,

metros y tranvías de una determinada agencia de transporte. La aplicación Moovit TimePro debe estar abierta en todo vehículo por el conductor, lo que permite conocer la ubicación y calcular de forma muy eficiente, los tiempos estimados de llegada, además de dar detalles al conductor sobre la ruta y si este ha llegado pronto, tarde o a tiempo. Los administradores de tráfico también pueden controlar y comparar en tiempo real estos vehículos (Azima, 2018).

Un ejemplo de negocio que se alimenta de todos estos conceptos; datos masivos, aprendizaje automático en la movilidad y soluciones MaaS es Uber. Uber AI realiza varias iniciativas al año, con el fin de mejorar el funcionamiento de la aplicación y la experiencia de usuario. El equipo de percepción mejora las capacidades de ubicación, cobertura, velocidad y direcciones de los vehículos, superando limitaciones del GPS convencional. El equipo de visión ha creado una identificación por imagen rápida y escalable para sus conductores, de tal forma que el ecosistema sea más seguro. Además, a través de modelos de aprendizaje automático de redes neurales mejoran su predicción de demanda total con el fin de prever problemas y estimar mejor sus tarifas, miden de mejor manera los tiempos estimados de llegadas y salidas de los conductores o desarrollan el taxi autónomo del futuro. Todo ello permite a Uber operar mejor que sus competidores (Ghahramani, 2019). Se conoce también que Uber utiliza la plataforma de Moovit en sus operaciones.

4. BiciMAD, las bicicletas eléctricas de Madrid

4.1 Introducción al caso y problemática

BiciMAD es uno de los sistemas de movilidad urbana más novedosos de Madrid. Se trata de un sistema de bicicletas eléctricas por alquiler de carácter público y gestionado, actualmente, por la EMT de Madrid. En la actualidad, el servicio dispone de más de 2.500 bicicletas y 200 estaciones operativas aproximadamente (BiciMAD, 2020).

Este caso práctico pretende enseñar la riqueza que pueden aportar los datos a BiciMAD y al sector de la movilidad urbana. El objetivo de este caso práctico es demonstrar cómo el análisis de datos ayuda en la mejora de la gestión y el conocimiento del negocio. Por ello, se realiza en primer lugar, un análisis exploratorio para conocer el comportamiento

de los usuarios y sus características demográficas y un análisis climatológico de la demanda del servicio, con el fin de ver cómo afectan variables ajenas a la demanda diaria de BiciMAD.

Posteriormente, se realiza un análisis de asimetría para conocer el desbalanceo que surge en el sistema, anteriormente comentado en los objetivos de este TFG, donde en algunas estaciones entran más bicicletas y de otras estaciones salen más bicicletas. Los empleados son aquellos que deben mover las bicicletas de estaciones que presentan un exceso a aquellas estaciones que presentan un defecto en el número de bicicletas disponibles, por lo que conoceremos las zonas con mayor exceso (asimetría positiva) y las zonas con mayor defecto (asimetría negativa).

Finalmente, utilizando el mes de septiembre de 2018, se realiza un análisis de "clustering" por K-means con variables de diversa índole (ocupación, número de reservas, capacidad de la estación etc.) con el fin de obtener grupos de estaciones a los cuales se les debe dar un tratamiento operativo diferente y otro análisis de "clustering" con series temporales donde se compara la media de ocupación (entendida como el número de bicicletas medio frente al número de bases/enganches operativos) de cada estación por hora para todo el mes de septiembre (24 mediciones por cada estación) con el fin de poder balancear mejor el sistema. Para este análisis, debemos recordar que BiciMAD está en continuo crecimiento y que muchas estaciones cambian de lugar, aumentan y disminuyen su número de bases, cambia el comportamiento de los usuarios etc. por lo que este estudio no es definitivo, pero sí es un primer acercamiento al análisis de BiciMAD y una propuesta a solucionar algunos de los problemas que sufren estos sistemas de movilidad sostenible en todas las ciudades.

4.2 Explicación de las bases de datos y variables utilizadas

4.2.1 Explicación de las bases de datos utilizadas

El estudio se compone de tres bases de datos:

1. Base de datos anual de uso que describe los trayectos realizados por los usuarios de BiciMAD en 2018. Es utilizada para un primer análisis exploratorio. 3,6 millones de observaciones o usuarios conforman esta base de datos. Los datos han sido descargados del portal OPENDATA de la EMT de Madrid.

2. Base de datos mensual de septiembre de 2018. Se compone de variables que miden el número de bicicletas y bases disponibles de cada estación para todos los días y rangos horarios de este mes. Es utilizada para realizar un análisis de "clustering" y "clustering" por series temporales. 172 observaciones o estaciones conforman esta base de datos. Los datos han sido descargados del portal OPENDATA de la EMT de Madrid.
3. Base de datos del clima de 2018. Se utilizan los datos de la AEMET o Agencia Estatal de Meteorología, medidos por su estación climática en el parque El Retiro. Esta base de datos contiene datos climáticos para cada día del año. Es utilizada para un análisis descriptivo de cómo el clima afecta a la demanda diaria. 365 observaciones o días conforman esta base de datos. Los datos han sido descargados de la página web datosclima.es, un portal que realiza "web scraping" para recoger los datos climáticos de la AEMET.

4.2.2 Explicación de las variables que contienen cada una de las bases de datos

La primera base de datos contiene los trayectos realizados por los usuarios anualmente que comprende, en la siguiente tabla, las variables que la conforman y que se puede encontrar en el portal OPENDATA de la EMT de Madrid (Portal OPENDATA - EMT, 2017). Únicamente se muestran las variables utilizadas.

Tabla 1. Descripción de Bases de datos 1: trayectos realizados en 2018

Nombre de la variable	**Descripción de la variable**
_id	Identificador del movimiento.
user_day_code	Código del usuario. Para una misma fecha, todos los movimientos de un mismo usuario tendrán el mismo código, con el fin de poder realizar estudios estadísticos de las tendencias diarias de los usuarios.
idunplug_station	Número de la estación de la que se desengancha la bicicleta.
idunplug_base	Número de la base o enganche de la que se desengancha la bicicleta.
idplug_station	Número de la estación en la que se engancha la bicicleta.
idplug_base	Número de la base o enganche en la que se engancha la bicicleta.

unplug_hourTime	Franja horaria en la que se realiza el desenganche de la bicicleta. Por cuestiones de anonimato, se facilita la hora de inicio del movimiento, sin la información de minutos y segundos. Todos los movimientos iniciados durante la misma hora tendrán el mismo dato de inicio.
travel_time	Tiempo total en segundos, entre el desenganche y el enganche de la bicicleta.
user_type	Número que indica el tipo de usuario que ha realizado el movimiento. Sus posibles valores son: 0: No se ha podido determinar el tipo de usuario 1: Usuario anual (poseedor de un pase anual) 2: Usuario ocasional 3: Trabajador de la empresa
ageRange	Número que indica el rango de edad del usuario que ha realizado el movimiento. 0: No se ha podido determinar el rango de edad del usuario 1: El usuario tiene entre 0 y 16 años 2: El usuario tiene entre 17 y 18 años 3: El usuario tiene entre 19 y 26 años 4: El usuario tiene entre 27 y 40 años 5: El usuario tiene entre 41 y 65 años 6: El usuario tiene 66 años o más
address_start	Variable añadida. Nombre de la calle de la estación de desenganche o “idunplug_station”.
long_start	Variable añadida. Coordenada de longitud de la estación de desenganche o “idunplug_station”.
lat_start	Variable añadida. Coordenada de latitud de la estación de desenganche o “idunplug_station”.
address_end	Variable añadida. Nombre de la calle de la estación de enganche o “idplug_station”.
long_end	Variable añadida. Coordenada de longitud de la estación de enganche o “idplug_station”.
lat_end	Variable añadida. Coordenada de latitud de la estación de enganche o “idplug_station”.
distance	Variable añadida. Distancia euclídea en línea recta desde la estación de desenganche y enganche. Creada a partir de las variables de longitud y latitud de estación de desenganche y enganche.
speed	Variable añadida. Distancia recta dividida por tiempo de recorrido o variable de “distance” dividida por la variable de “travel_time”.

La segunda base de datos es el estado de las estaciones de septiembre de 2018. Esta base de datos tiene catorce variables idénticas para cada estación que a continuación se describen en la siguiente tabla.

Tabla 2. Descripción de Bases de datos 2: Estado de las estaciones durante septiembre 2018

Nombre de la variable	**Descripción de la variable**
_id	Fechas y horas en la que se mide el estado de cada estación "x" de BiciMAD.
stations.x.activate	Variable que mide si una estación "x" está realizando mediciones sobre su estado o no. 1: activada 0: no activada
stations.x.address	Muestra la dirección completa de la estación "x".
stations.x.dock_bikes	Variable que mide el número de bicicletas ancladas en la estación "x" en ese momento.
stations.x.free_bases	Variable que mide el número de bases disponibles en la estación "x" en cada momento. Este número suele ser menor que el expresado en la variable stations.x.total_bases, ya que las estaciones suelen tener bases inoperativos o dañados.
stations.x.id	Indica el identificador para cada estación "x".
stations.x.latitude	Indica la coordenada de latitud para cada estación "x". Muy útil para la elaboración de mapas.
stations.x.longitude	Indica la coordenada de longitud para cada estación "x". Muy útil para la elaboración de mapas.
stations.x.name	Muestra el nombre que recibe la estación "x" por el sistema de BiciMAD.
stations.x.number	Muestra el número que recibe la estación "x" por el sistema de BiciMAD.
stations.x.reservations_count	Indica el número de reservas que recibe la estación "x" en cada momento.
stations.x.total_bases	Indica el número total de bases para cada estación "x", independientemente de si estos están dañados o inoperativos.

Para realizar el análisis de "clustering" se ha transpuesto esta base de datos: en vez de tener catorce columnas/variables para cada estación y los mismos registros de tiempo y fecha para todas, se ha optado por tener catorce variables para todas las estaciones, creando una variable con el nombre de la estación y tener los mismos registros de tiempo duplicados para cada estación.

Por último, la tercera se refiere a los datos climáticos de 2018, con datos de la AEMET, medidos por la estación meteorológica del parque de El Retiro y contiene una medición diaria de las siguientes variables climáticas:

Tabla 3. Descripción de Bases de datos 3: Datos climáticos diarios 2018

Nombre de la variable	**Descripción de la variable**
Date	Muestra la fecha en formato (Año-Mes-Día)
Station_id	Muestra el identificador de la estación meteorológica, en este caso siempre será el parque de El Retiro
Tmean	Indica la temperatura media (ºC).
Tmax	Indica la temperatura máxima (ºC).
Tmin	Indica la Temperatura mínima (ºC).
Rain_amount	Indica la precipitación o cantidad de lluvia (mm = l/m2).
Wind_Direction	Indica la dirección del viento (decenas de grado).
Wind_speed	Indica la velocidad media (m/s).
Wind_Gust	Indica la racha de viento máxima alcanzada (m/s).
Pressure_max	Indica la máxima presión atmosférica alcanzada (hPa).
Pressure_min	Indica la mínima presión atmosférica alcanzada (hPa).

Una vez descritas las variables de las bases de datos utilizadas, se procede a enseñar los distintos análisis realizados del sistema de la bicicleta eléctrica de Madrid, BiciMAD.

4.3 Análisis exploratorio de los trayectos realizados en el año 2018

En este primer análisis exploratorio, el objetivo es el de comprender cómo es el comportamiento de los usuarios de BiciMAD a lo largo de 2018 y sirve para dar una visión holística de los trayectos/viajes realizados, además de contabilizar el número de usuarios que hay para una determinada variable o un conjunto de estas. Si se desea mayor nivel de detalle, se podría, de la misma manera, explorar el comportamiento por cada estación. Este análisis tiene un fin descriptivo, ya que no se pueden extrapolar estos comportamientos a otros años como 2017 o 2019. Para lograr este cometido, se han realizado una serie de gráficas, a modo de informe visual, que podrán ver en las siguientes páginas de este documento.

4.3.1 Análisis exploratorio de la variable de grupos/intervalos de edad

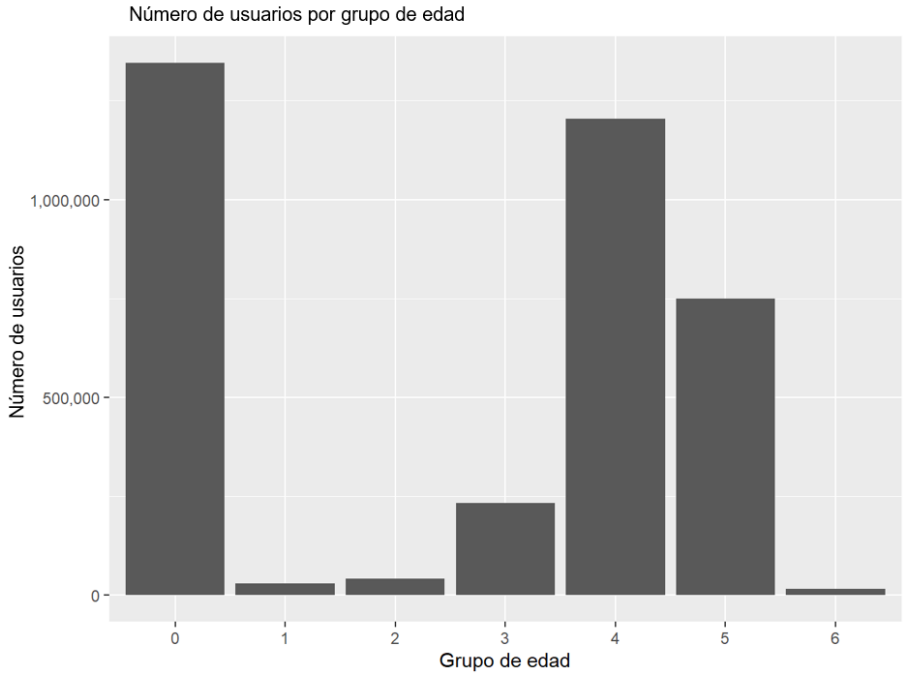


Ilustración 9. BiciMAD - Número de usuarios por grupo de edad (Elaboración propia).

La primera gráfica, ilustración 9, surge de la pregunta ¿Quiénes utilizan más el servicio? Como se ha explicado anteriormente, en las variables de cada base de datos, el grupo de edad o ageRange está dividida en seis intervalos de años. Los grupos con mayor número de viajes realizados son el cero, cuatro, cinco y tres. Esta gráfica da a entender que a la mayoría de los usuarios del sistema no se les puede identificar ya que hay 1,4 millones de viajes en esta categoría (grupo cero). Dentro de los grupos más numerosos, los usuarios que tienen entre 27 y 40 años (grupo cuatro) son los que conforman el grueso en número de viajes realizados, con alrededor de 1,2 millones viajes, seguido de los usuarios que tienen entre 41 y 65 años (grupo cinco) con 750 mil viajes y, por último, seguido por aquellos usuarios entre 19 y 26 años (grupo tres) con alrededor de 240 mil viajes. Se debe tener en cuenta que la variable ageRange, agrupa intervalos más largos y cortos de años. Por ejemplo, el grupo cuatro agrupa trece años, el grupo cinco agrupa veinticuatro y el grupo dos, tan solo dos. Los grupos con pocos viajes al año son aquellos usuarios desde los 0 a los 16 años (grupo uno), desde los 17 a los 18 años (grupo dos) y con más de 66 años (grupo seis).

Una vez comprendidos el número de viajes efectuados por rango de edad, se decide ver el comportamiento de las edades respecto a la variable de distancia. Esta variable de distancia es calculada dibujando una línea recta entre dos puntos, en este caso, estación

de origen y estación de destino, siguiendo los principios de la distancia euclidiana. Por lo tanto, la variable presupone el comportamiento de los usuarios como el más eficiente y que estos escogerán la ruta más rápida y en línea recta. Esta suposición es arriesgada y probablemente no sea cierta, pero es válida para este análisis, ya que los usuarios suelen tardar más o menos dependiendo de cómo de alejadas estén las estaciones de origen y destino, sin importar el método para calcular la distancia.

En la ilustración 10, se pueden observar varios comportamientos. El intervalo comprendido entre 0 y 16 años (grupo uno) presenta su mayor número de casos sobre una distancia más elevada, 1,9 kilómetros aproximadamente. Además, se puede observar que estos viajes presentan una distancia mayor en general, cuando se comparan con el resto de los grupos. Si nos vamos al otro extremo, conformado por aquellos usuarios con más de 66 años (grupo seis) podremos ver que los viajes situados en una distancia de 1 kilómetro o inferior son muy superiores al resto de grupos. Los demás intervalos de edad presentan características muy parecidas.

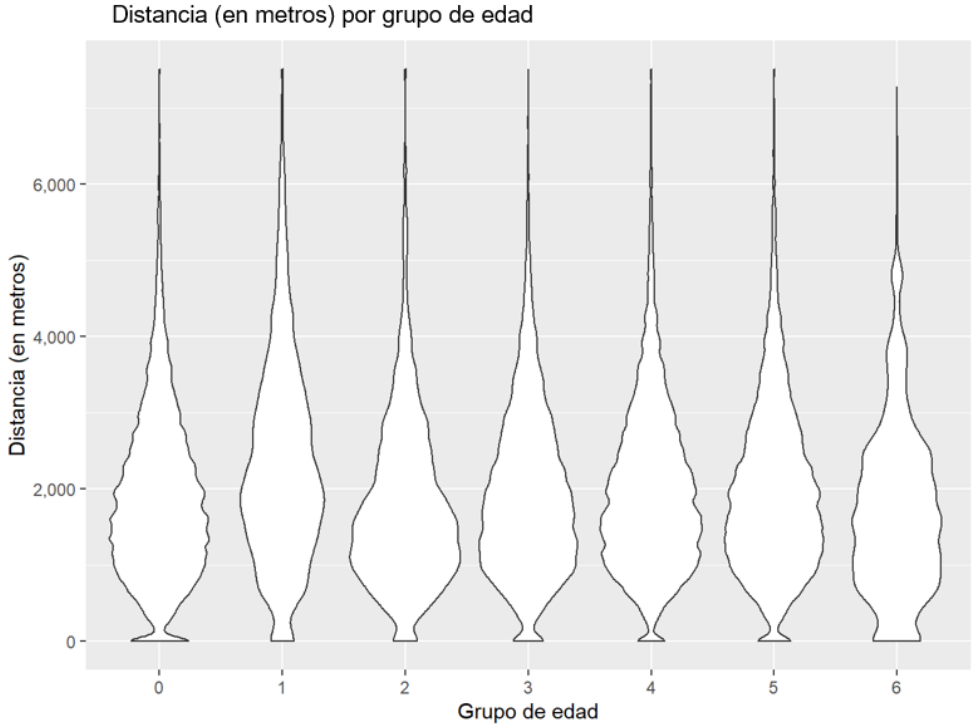


Ilustración 10. BiciMAD – Distancia por grupo de edad (Elaboración propia).

4.3.2 Análisis exploratorio de la variable tipo de usuario

A continuación, se pasa a analizar el comportamiento de los usuarios respecto al tipo al que corresponden. Este conjunto de gráficos surge de la pregunta ¿Difiere la duración por recorrido de los usuarios según su tipo?

La primera gráfica, ilustración 11, mide la duración del trayecto en segundos respecto al tipo de usuario. El tipo de usuario cero es aquel al que no se ha podido identificar, el tipo uno son aquellos usuarios que poseen un pase/abono anual, el tipo dos son usuarios ocasionales y el tipo tres, son empleados del servicio. Como se puede observar, el comportamiento difiere bastante por tipo de usuario.

La duración del tipo de usuario uno, que normalmente utiliza la bicicleta durante la semana y a la hora de trabajar, es muy inferior a la del tipo de usuario dos, que posiblemente quiera sacar el mayor partido a la bicicleta en vacaciones, fines de semana, días festivos etc. Si se presta atención a ambos grupos, se puede observar que el tipo de usuario uno expresa unos cuartiles más equidistantes a la media que el tipo de usuario dos. Esto puede significar que el usuario del tipo uno se comporta de una manera más homogénea que aquellos del grupo dos. Por último, el tipo de usuario tres son empleados y es totalmente lógico que tengan la media más elevada ya que desenganchan bicicletas e intentan balancear el sistema de la mejor manera posible, pudiendo ir hasta estaciones muy alejadas en comparación con el tipo de usuario uno y dos. La gráfica ha sido acotada hasta los 7500 segundos o 2 horas debido a la presencia de valores atípicos.

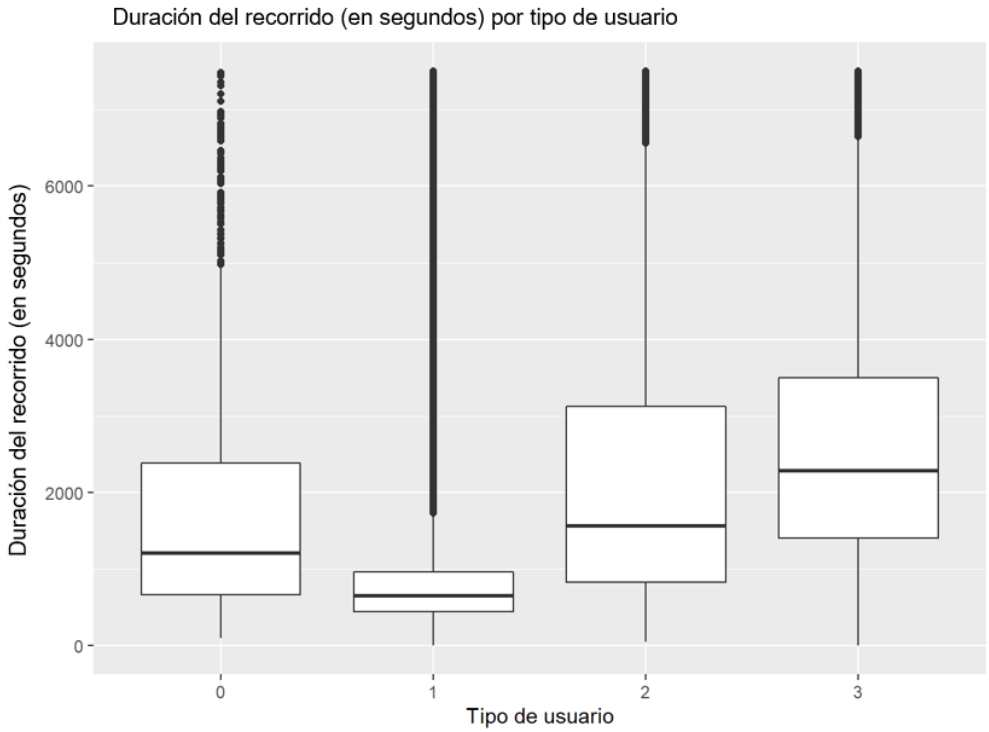


Ilustración 11. BiciMAD – Duración del recorrido por tipo de usuario (Elaboración propia).

La siguiente gráfica, ilustración 12, surge de lo hallado en la anterior, ¿Realmente, la diferencia en la duración del recorrido por tipo de usuario surge de la realización de distancias más largas?

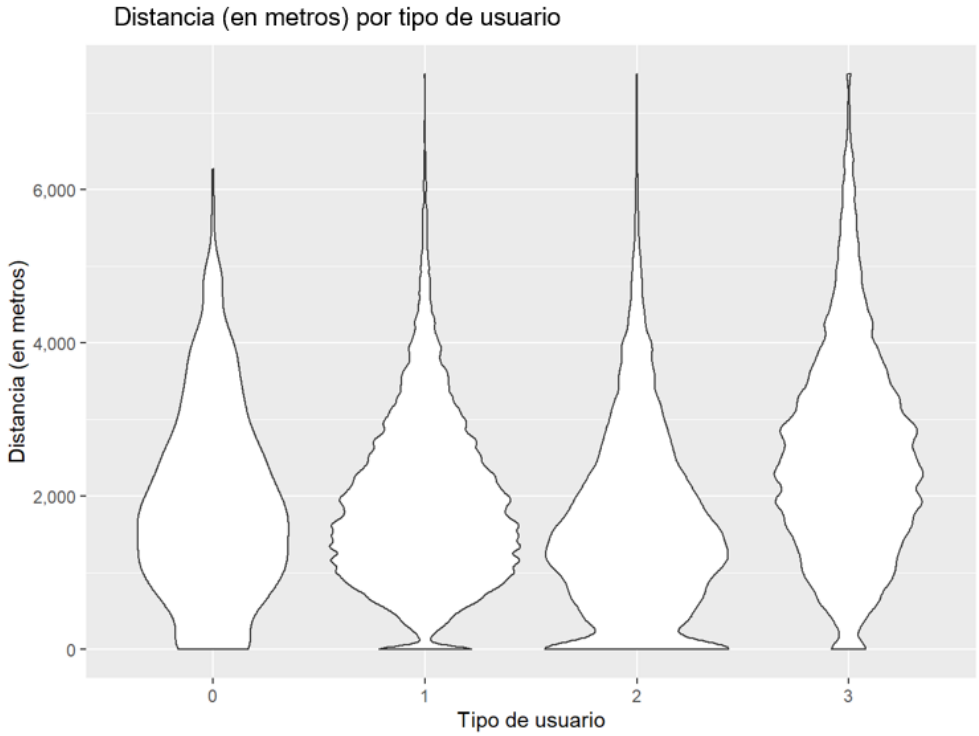


Ilustración 12. BiciMAD – Distancia por tipo de usuario (Elaboración propia).

Nótese que, con las representaciones gráficas de estos datos, no solamente buscamos el describir los datos sino entender las variables y establecer relaciones entre sí. Esta gráfica da a entender que como anteriormente se ha expuesto, la mayor duración por viaje de los empleados o tipo de usuario tres viene dada por las mayores distancias recorridas. No obstante, no se da el caso para el tipo de usuario uno y dos, que exhiben un comportamiento casi idéntico, aunque se concentra un mayor número de viajes en distancias inferiores a un kilómetro en el tipo de usuario dos que en el tipo de usuario uno. También se pretende conocer cuál es la velocidad media en estos 3,6 millones de viajes por tipo de usuario, para comprender cómo se efectúan los viajes y saber quién es el grupo que efectúa los viajes a mayor velocidad.

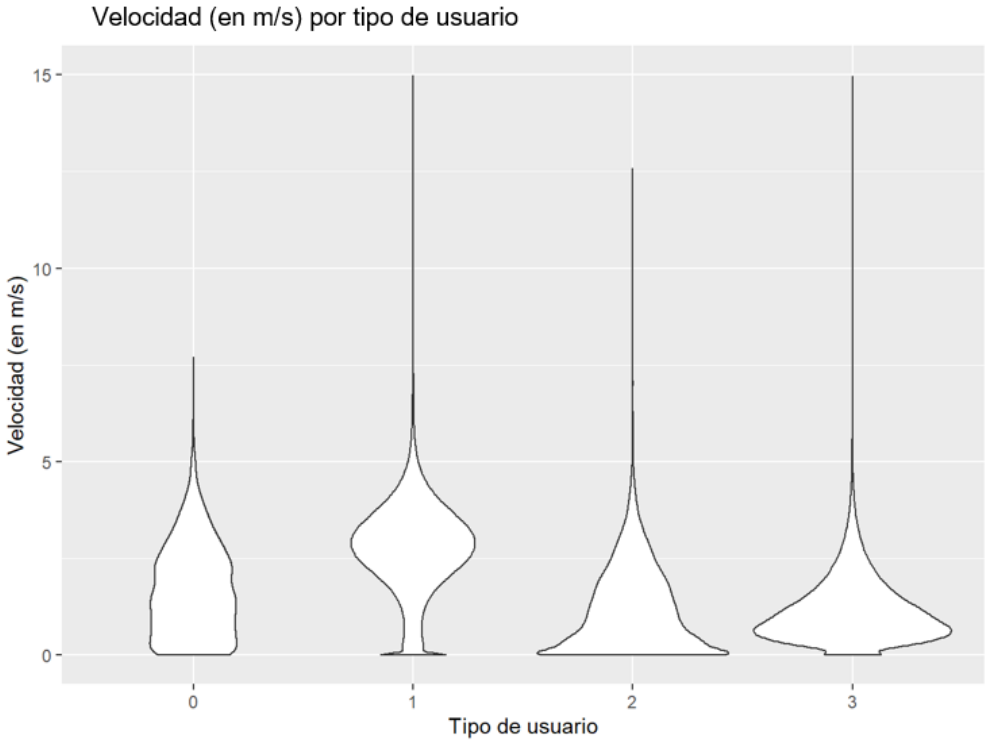


Ilustración 13. BiciMAD – Velocidad por tipo de usuario (Elaboración propia).

En esta gráfica se puede determinar que el tipo de usuario uno es el más veloz, seguido del tres y por último del dos. La variable de velocidad se ha calculado dividiendo la variable de "Distance" entre "travel_time". Recordemos que la variable de "Distance" o distancia está sujeta a algunas condiciones (distancia euclídea y máxima eficiencia), por lo que la variable de velocidad también está sujeta a estas condiciones. Aun así, se puede comprobar de manera genérica que el tipo de usuario uno realiza trayectos más veloces, probablemente por razones de trabajo, mientras que el tipo de usuario dos tiene la velocidad más baja, ya que su trayecto se considera una actividad lúdica.

Se ha relacionado, en la ilustración 14, el número de viajes por tipo de usuario y por el grupo de edad al que pertenece con una muestra en un mapa de calor. Visualizamos al tipo de usuario uno como el más numeroso, con colores más claros en casi todos los rangos de edad. Por lo que se puede ver, el tipo de usuario dos no es muy numeroso ni está muy identificado, con la mayoría de los casos cayendo al grupo de edad cero. No se han clasificado los tipos de usuario 0 (debido a ser un usuario no identificado) ni 3 (ya que son empleados y no usuarios del sistema).

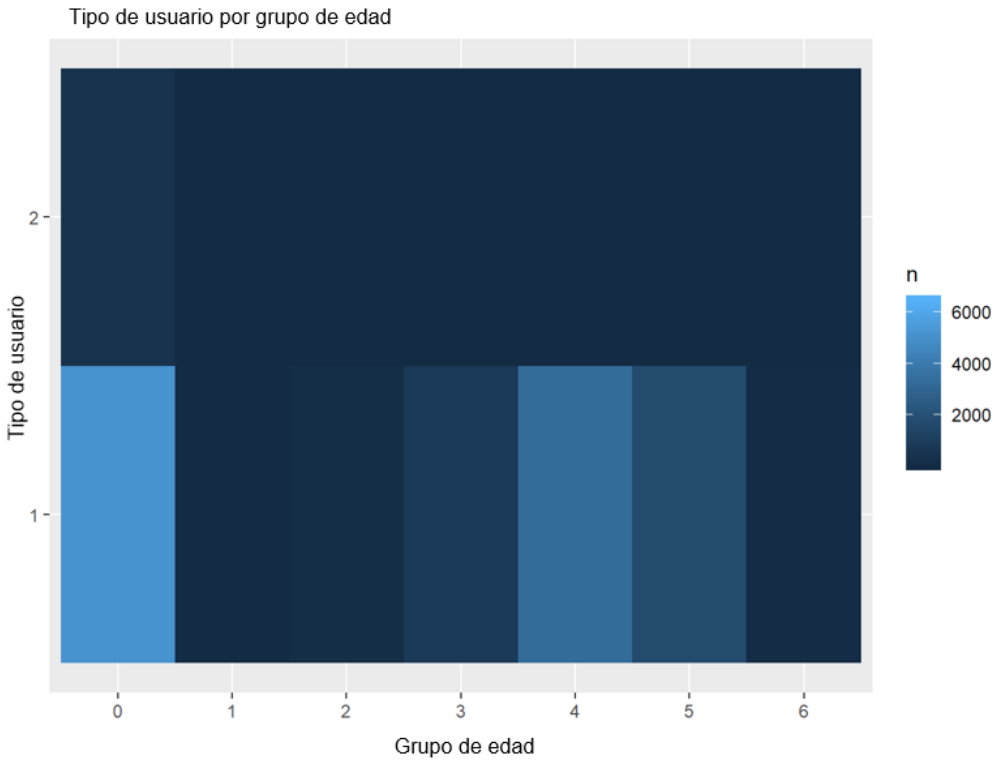


Ilustración 14. *BiciMAD – Número de viajes por grupo de edad y tipo de usuario (Elaboración propia).*

4.3.3 Análisis exploratorio de día de la semana, franjas horarias, y estación del año

Al tener variables como la fecha y hora del viaje, es inevitable no pensar en el posible cambio de comportamiento del usuario según estas variables de carácter temporal. Esta primera gráfica o ilustración 15, describe el número de viajes en 2018 por día de la semana; donde lo más característico es el menor nivel de viajes en el fin de semana frente a los días laborales. Además, se visualiza que el lunes y el viernes, días más cercanos a sábado y domingo, hay menos viajes que el resto de la semana, siendo el martes, miércoles y jueves los días con más viajes, con registros cercanos a los 590 mil viajes.

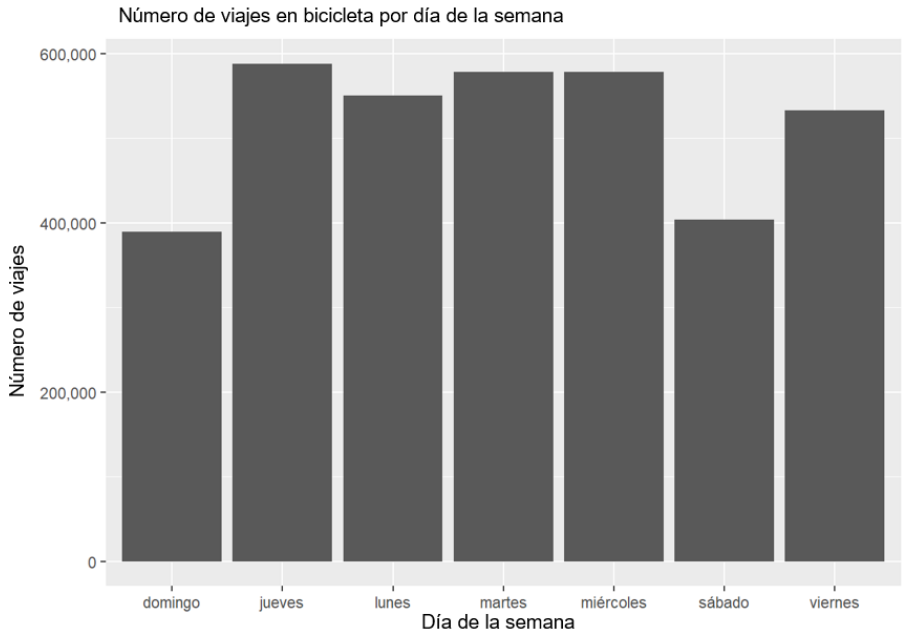


Ilustración 15. BiciMAD – Número de viajes por día de la semana (Elaboración propia).

La siguiente pregunta, algo más elaborada, es si hay diferencia en la edad respecto al día de la semana. Se han utilizado porcentajes para poder visualizar mejor la ilustración 16.

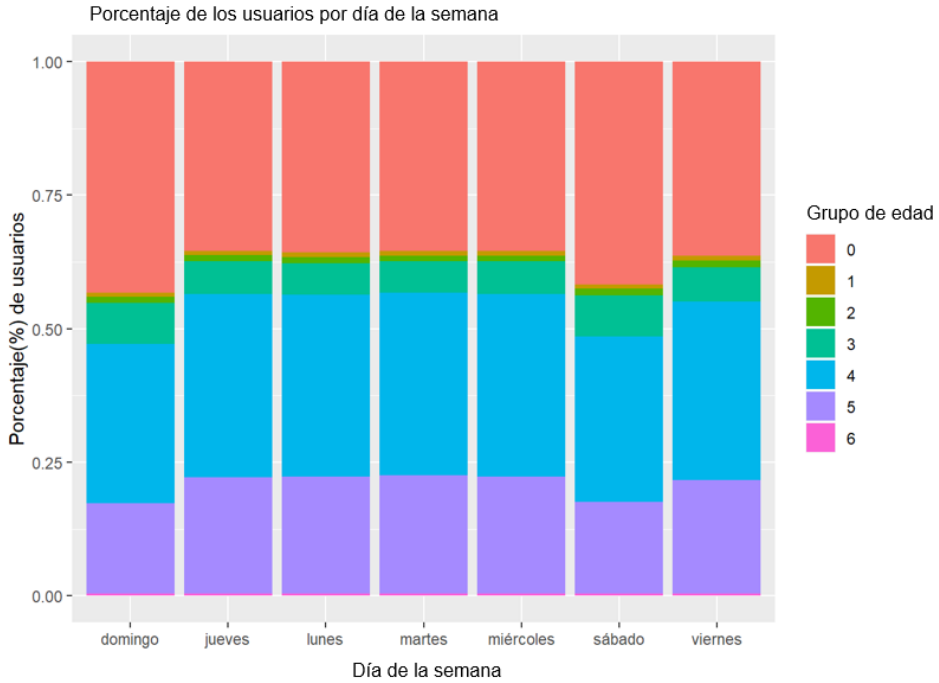


Ilustración 16. BiciMAD – Porcentaje de los usuarios por día de la semana y grupo de edad (Elaboración propia).

En la ilustración 16, se puede comprobar que el fin de semana hay más usuarios cero o aquellos usuarios a los que no se puede identificar su grupo de edad. En el anterior mapa de calor, ilustración 14, que relacionaba la cantidad de usuarios por tipo de usuario y por grupo de edad, se puede visualizar que la mayoría de los usuarios con tipo de usuario dos

están clasificados con un grupo de edad cero. Teniendo en cuenta que hay más usuarios con grupo de edad cero en los fines de semana, este gráfico podría significar que hay más usuarios ocasionales o tipo de usuario dos en estos días.

La siguiente gráfica muestra cómo se han comportado los usuarios de BiciMAD a lo largo de las estaciones del año. Las estaciones se han separado por solsticios y equinoccios del año, teniendo en cuenta la fecha por viaje efectuado. Este gráfico enseña la duración del recorrido según la estación del año.

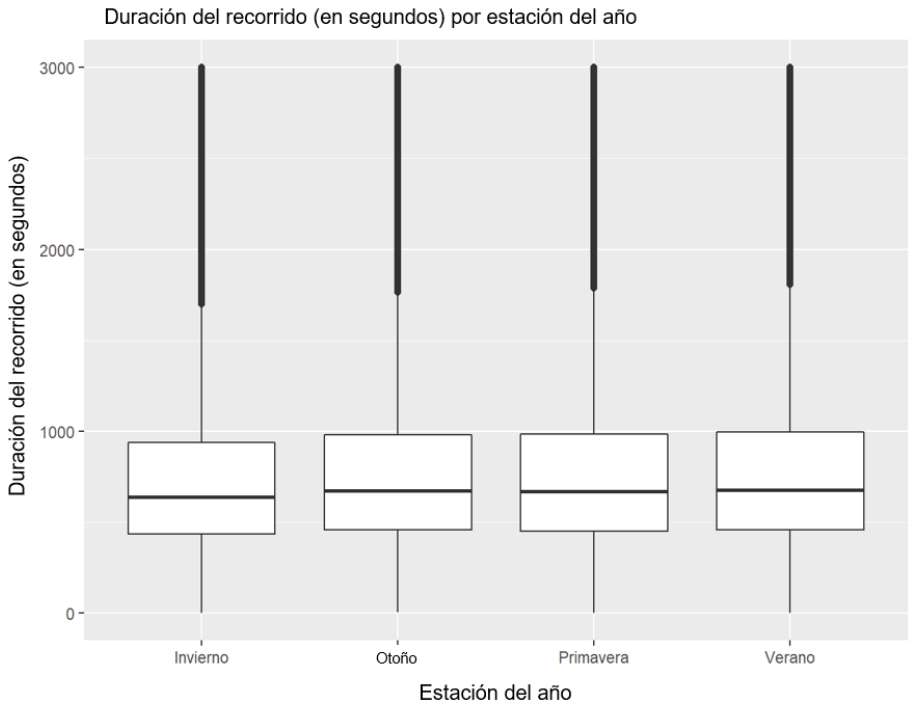


Ilustración 17. BiciMAD – Duración del recorrido por estación del año (Elaboración propia).

En invierno, se muestra una duración media menor, probablemente debido a la menor temperatura y peor tiempo. En otoño se registra la mayor media de duración y sus cuartiles son más elevados respecto a otras estaciones. Esto se puede deber a que no hay tanta precipitación como en primavera ni tanto calor como en verano, que registran medias algo inferiores. Aun así, la duración es bastante similar a lo largo de todo el año.

También se ha creado una variable de franja horaria, teniendo en cuenta las horas a las que se efectúa cada viaje. El objetivo es conocer que tipos de usuarios utilizan más el servicio y en que franja horaria. Los resultados de la ilustración 18 son los siguientes; el tipo de usuario que predomina es el uno o abonado anual y utiliza el servicio por la mañana y tarde, probablemente para ir a trabajar por la mañana y volver del trabajo por la tarde. Los usuarios de tipo dos u ocasionales son más numerosos por la mañana y tarde-

noche. Por último, también podemos conocer los hábitos de los usuarios de tipo tres o empleados que realizan el rebalanceo del sistema en las franjas horarias de tarde-noche y noche.

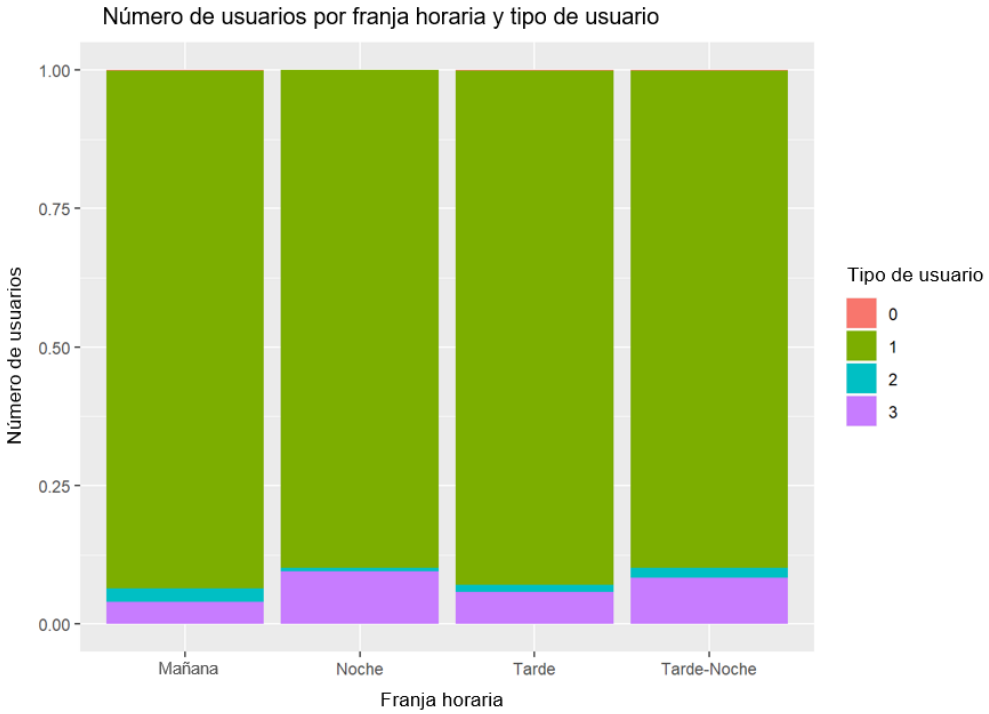


Ilustración 18. BiciMAD – Número de usuarios por franja horaria y tipo de usuario. (Elaboración propia).

4.4 Análisis del efecto del clima en la demanda del año 2018

A continuación, se realiza un análisis descriptivo de la demanda agregada diaria frente a variables climáticas. Este análisis nos sirve para observar cómo variables que no están a nuestro alcance, como el clima, pueden afectar en el uso del servicio y concretamente, para realizar previsiones de demanda diaria con el fin de planificar el número de bicicletas expuestas al público. Se ha de mencionar que, aunque el servicio BiciMAD esté operativo las 24 horas del día, es utilizado mayoritariamente por el día. Al tener datos climáticos diarios y no horarios, esto puede dar lugar a limitaciones en este análisis donde, por ejemplo, la demanda diaria es alta durante el día y las precipitaciones son altas durante la noche cuando no se utilizaba el servicio, causando una incorrecta interpretación de los datos.

La demanda agregada diaria es calculada con la base de datos 1 de 2018 explicada en el apartado 4.2.1, donde se agrupa la cantidad de viajes por cada día de servicio del sistema. Las variables climáticas las contiene la base de datos 3, que recoge mediciones climáticas

importantes de la AEMET a lo largo de los 365 días del año 2018. Además, se han elegido los datos medidos por la estación meteorológica Parque El Retiro, puesto que es una de las zonas más próximas a la zona de mayor actividad de BiciMAD, el Centro histórico.

4.4.1 Relaciones entre la demanda diaria y el clima: temperatura (centígrados)

En estas primeras gráficas, ilustraciones 19 y 20, se comparan las variables de temperatura media y temperatura máxima frente a la demanda diaria. La relación no es lineal, realizando una parábola. La demanda diaria crece hasta los 20 grados centígrados aproximadamente de temperatura media y decrece en torno a los 22-23, posiblemente debido al calor. La temperatura máxima es más permisiva, donde la demanda diaria aumenta hasta los 25 grados centígrados, punto en el cual la demanda empieza a decrecer.

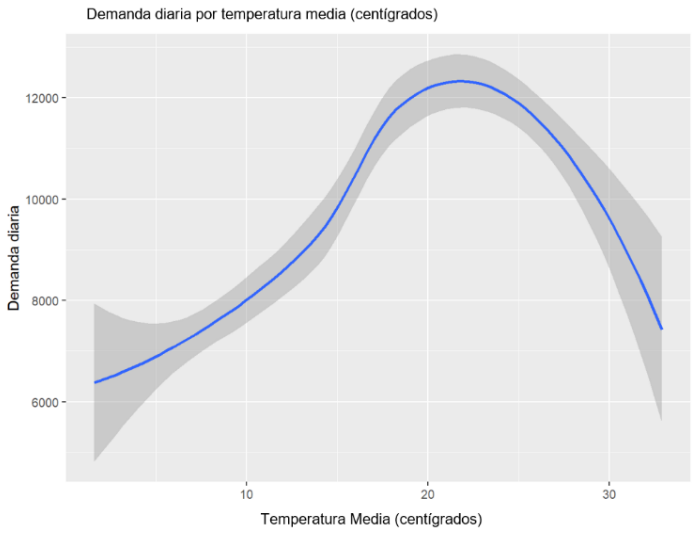


Ilustración 19. BiciMAD – Demanda diaria por temperatura media diaria (Elaboración propia).

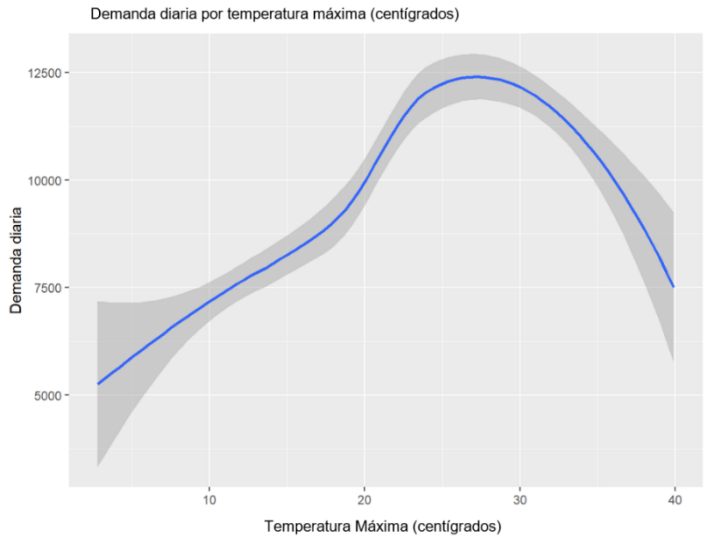


Ilustración 20. BiciMAD – Demanda diaria por temperatura máxima diaria (Elaboración propia).

4.4.2 Relaciones entre la demanda diaria y el clima: viento (m/s)

A continuación, se explican las variables relacionadas del viento. Esta relación en ambos casos no es lineal. Se aprecia un aumento de la demanda debido a la velocidad y racha del viento que eventualmente decrece. Los usuarios dejan de coger las bicicletas cuando hay bastante viento, y cuando este se produce a grandes velocidades. Estas graficas son útiles debido a que el viento a velocidades elevadas puede causar inestabilidad e inseguridad en la bicicleta derivando en mayor número de accidentes.

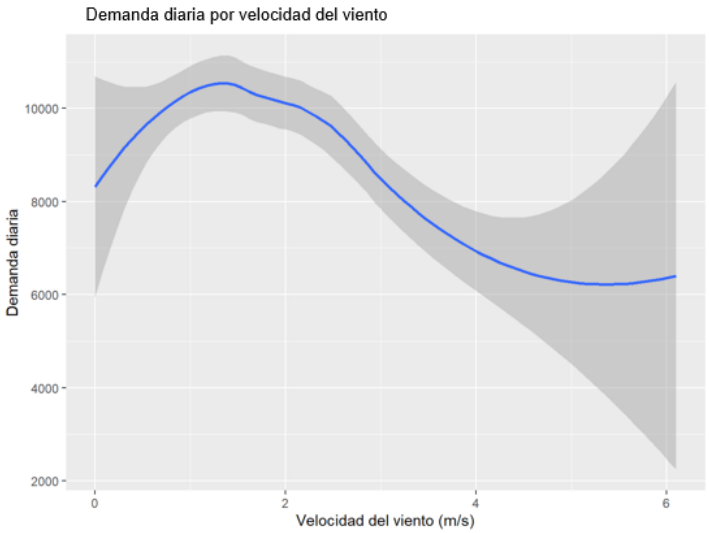


Ilustración 21. BiciMAD – Demanda diaria por velocidad del viento diaria (Elaboración propia).

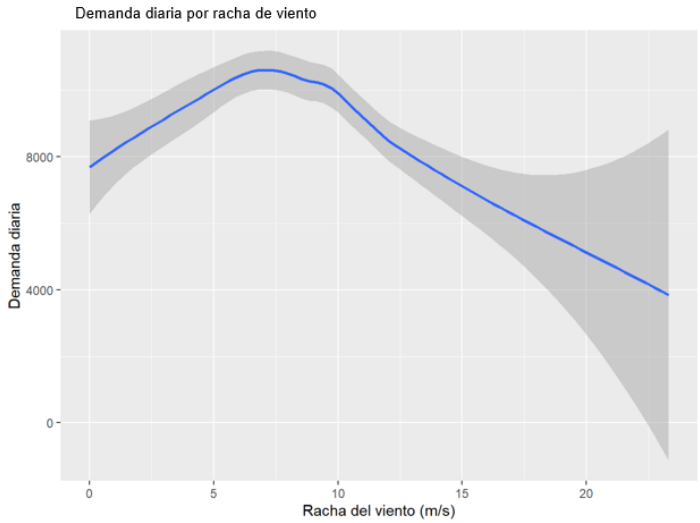


Ilustración 22. BiciMAD – Demanda diaria por racha de viento diaria (Elaboración propia).

4.4.3 Relaciones entre la demanda diaria y el clima: precipitación (mm=l/m2)

En este gráfico, ilustración 23, la demanda diaria cae a medida que aumenta la lluvia hasta los 16 litros por metro cuadrado. No obstante, existe un punto de inflexión donde los niveles de lluvia más altos aumentan la demanda, tal vez porque haya llovido a una determinada hora donde los niveles de demanda no fueran muy altos como, por ejemplo, que haya llovido fuertemente por la noche. Recordemos que únicamente se dispone de datos diarios y no se distingue por franja horaria.

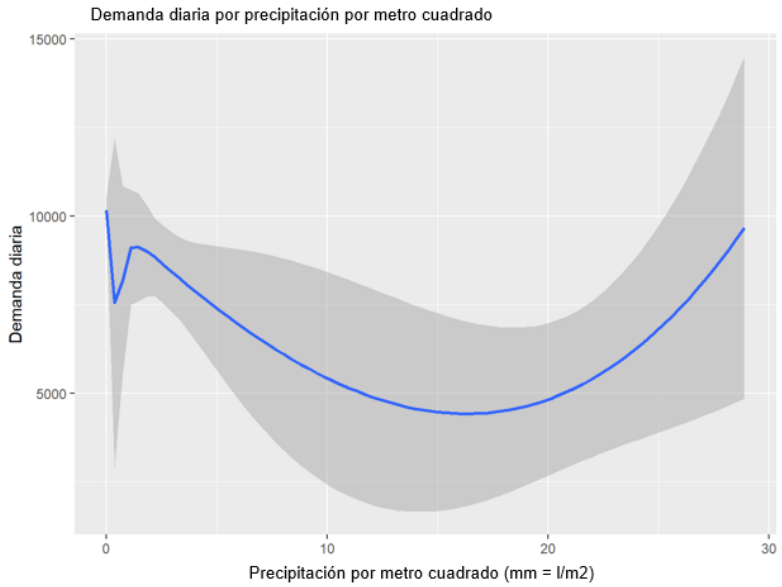


Ilustración 23. BiciMAD – Demanda diaria por precipitación por metro cuadrado (Elaboración propia).

4.5 Análisis de asimetría entre estaciones y rutas más populares

Una vez comprendida la base de datos 1 y el comportamiento de las observaciones, con un análisis exploratorio y del clima, se debe explorar y medir el problema más habitual que tienen estos sistemas de bicicletas: su asimetría. Este análisis está basado en el estudio realizado por la universidad de Cornell "Data Analysis and Optimization for (Citi)Bike Sharing" en 2015 donde se explica el concepto de asimetría (O'Mahony & Shmoys, 2015).

Se ha denominado asimetría a la diferencia de bicicletas que se enganchan y se desenganchan de una estación determinada, de tal forma que:

Asimetría de X estación = Nº llegadas a estación X – Nº Salidas de estación X

Este problema es frecuente ya que los usuarios tienden a enganchar las bicicletas en un tipo determinado de estaciones (Nºllegadas > Nºsalidas) y a desengancharlas en otro tipo determinado de estaciones (Nºllegadas < Nºsalidas), creando un fuerte desbalanceo que es compensado mediante los empleados. Estos empleados mueven estas bicicletas de aquellas estaciones con un número de llegadas muy alto a aquellas con un número de llegadas muy bajo, de tal forma que ninguna estación se quede sin bicicletas.

Con la base de datos de los trayectos o viajes realizados en 2018, se puede identificar, para cada viaje, la estación de origen y destino. Por consiguiente, se puede obtener la asimetría que hubo en las 172 estaciones operativas para un determinado periodo de tiempo a lo largo del año.

En primer lugar, se ha decidido calcular la asimetría por estaciones del año excluyendo los viajes efectuados por los empleados, para ver si hay alguna diferencia de carácter temporal y estacional, además de demostrar la labor que realizan estos últimos. Los datos resultantes se han ordenado por las diez primeras estaciones con más asimetría positiva (Nºllegadas > Nºsalidas) y por las diez primeras estaciones con más asimetría negativa (Nºllegadas < Nºsalidas) con el fin de determinar de mejor manera cómo evoluciona la asimetría de las estaciones a lo largo del año. Posteriormente, se ha analizado la asimetría total para cada estación del sistema, teniendo en cuenta e incluyendo los viajes efectuados por los empleados para balancear el sistema a lo largo de todo el año.

Es importante mencionar que los empleados efectúan el rebalanceo mediante furgonetas, donde cargan un gran número de bicicletas, por lo que los registros de la base de datos 1 no son viajes individuales o en bicicleta como tal, al contrario de lo que encontramos con los usuarios con abono anual y ocasional. En una última instancia y a modo de ejemplo, comprobaremos para una estación en concreto cual ha sido el desempeño de los empleados.

4.5.1 Análisis de asimetría positiva entre las estaciones del año

Tabla 4. Evolución de las estaciones con mayor asimetría positiva

	Asimetría positiva			
Nombre de la estación	**Invierno**	**Primavera**	**Verano**	**Otoño**
Paseo de la Chopera nº14	2.579 (1)	3.702 (1)	4.610 (1)	4.498 (1)
Paseo de la Florida nº8	2.230 (2)	2.868 (2)	2.706 (2)	2.723 (2)
C/Cerro de la Plata nº2	1.490 (3)	1.849 (3)	2.500 (3)	1.986 (4)
Paseo Santa María Cabeza nº58	1.239 (4)	1.513 (4)	1.853 (4)	2.206 (3)
Plaza de la Cebada nº16	963 (5)	1.203 (6)	1.068 (7)	1.202 (6)
Paseo de los Olmos nº28	899 (6)	1.328 (5)	1.392 (5)	1.081 (7)
C/Jaime el Conquistador nº30	856 (7)	794 (8)	1.267 (6)	1.294 (5)
Paseo de las Delicias nº92-94	636 (8)	896 (7)	1.053 (8)	780 (9)
C/Fernando el Católico nº19	629 (9)	650 (10)	694 (9)	876 (8)
C/Palos de la Frontera nº40	616 (10)	n.d	n.d	n.d
Avda. de Menéndez Pelayo nº38	n.d	700 (9)	n.d	581 (10)
Plaza de San Francisco nº5	n.d	n.d	690 (10)	n.d

(x) indica la posición de la estación dentro de las 10 primeras posiciones. n.d: la estación no se encuentra en las 10 primeras posiciones.

En esta primera tabla se resumen las estaciones con mayor número de asimetría positiva, esto es, aquellas estaciones con mayor número de llegadas frente a su número de salidas. Se debe tener en cuenta, que esta tabla no indica aquellas estaciones donde más bicicletas llegan en términos absolutos, sino en relativos, al compararlos con su número de salidas. Como se puede observar, las estaciones de Paseo de la Chopera nº14, Paseo de la Florida nº8, Calle Cerro de la Plata nº2 y Paseo de Santa María de la Cabeza nº58 son las que mayor número tienen. Se debe mencionar que no hay diferencias notables hasta el séptimo puesto en este análisis, lo que significa que las estaciones de llegada suelen ser las mismas sin importar demasiado la estación meteorológica del año. Los empleados deben, en la medida de lo posible, utilizar estas estaciones para realizar el rebalanceo el sistema desplazando bicicletas desde aquí a donde haya un defecto de estas, ya que, a mayor número de asimetría positiva, mayor número de bicicletas enganchadas en la estación.

Esta tabla también indica la posición de las estaciones y cómo estás ganan o pierden importancia respecto a la estación meteorológica.

4.5.2 Análisis de asimetría negativa entre las estaciones del año

Tabla 5. Evolución de las estaciones con mayor asimetría negativa

	Asimetría negativa			
Nombre de la estación	**Invierno**	**Primavera**	**Verano**	**Otoño**
Paseo Castellana nº67	-861 (1)	-1.017 (1)	-1.300 (1)	-1.072 (1)
Paseo de la Castellana nº4	-753 (2)	-944 (2)	-1.051 (2)	-950 (4)
C/Carlos III nº1	-607 (3)	-691 (5)	-900 (4)	-771 (5)
C/María de Guzmán nº58	-571 (4)	-595 (9)	n.d	-691 (9)
C/Valencia nº1	-534 (5)	n.d	n.d	n.d
C/Sor Ángela de la Cruz nº2	-509 (6)	n.d	-703 (9)	-724 (7)
Paseo Yeserías nº45	-506 (7)	n.d	n.d	n.d
Avda. de Menéndez Pelayo nº3	-495 (8)	-722 (4)	-940 (3)	-995 (3)
Puerta del Sol nº1 - B	-493 (9)	-578 (10)	-787 (6)	n.d
C/Serrano nº34 - A	-463 (10)	-727 (3)	-792 (5)	-1002 (2)
Plaza San Juan de la Cruz nº11	n.d	-681 (6)	n.d	n.d
Paseo Castellana nº42	n.d	-658 (7)	-724 (8)	n.d
C/Goya nº18	n.d	-610 (8)	n.d	n.d
C/ Serrano nº34 - B	n.d	n.d	-766 (7)	-731 (6)
C/Serrano nº54	n.d	n.d	-683 (10)	n.d
C/Pavía nº6	n.d	n.d	n.d	-701 (8)
Plaza de España - A	n.d	n.d	n.d	-594 (10)

(x) indica la posición de la estación dentro de las 10 primeras posiciones. n.d: la estación no se encuentra en las 10 primeras posiciones.

En esta segunda tabla se resumen las estaciones con mayor número de asimetría negativa, esto es, aquellas estaciones con mayor número de salidas frente a su número de llegadas. Se debe tener en cuenta, que esta tabla no indica aquellas estaciones donde más bicicletas salen en términos absolutos, sino en relativos, al compararlos con su número de llegadas.

Como se puede observar, las primeras estaciones como Paseo de la Castellana nº67, Paseo de la Castellana nº4, Calle Carlos III nº1, Avenida de Menéndez Pelayo nº3 y Calle Serrano nº34 son las que mayor número tienen. Se deduce también que los usuarios del sistema de BiciMAD inician su viaje de un número más amplio de estaciones al que lo finalizan, al haber mayor diversidad de estaciones en las diez primeras posiciones de asimetría negativa. También se puede visualizar que las estaciones de BiciMAD que aparecen varias veces en las diez primeras posiciones experimentan mayores subidas y bajadas en su posición que en la tabla anterior, donde se clasificaba aquellas con mayor asimetría positiva. Esto significa que hay diferencias de carácter temporal en el uso de las estaciones de origen o salida.

4.5.3 Análisis de asimetría total y anual

En este análisis se han incluido los usuarios de tipo 3 (empleados), para contabilizar el efecto que tiene el rebalanceo del sistema. En las siguientes tablas, se pueden observar las estaciones que más sufren de la poca o demasiada disponibilidad de bicicletas, incluso después del rebalanceo realizado por los empleados.

Tabla 6. Estaciones con mayor número de asimetría positiva total contabilizando empleados

Nombre de la estación	**Número de salidas efectuadas/desenganches**	**Número de llegadas efectuadas/enganches**	**Asimetría total**
Paseo de la Chopera nº14	51.926	53.274	1348
Paseo de la Florida nº8	34.844	35.572	728
Paseo de Santa María Cabeza nº58	48.710	49.370	660
Plaza de la Cebada nº16	50.067	50.650	583
C/Cerro de la Plata nº2	26.279	26.831	552
Plaza de Jacinto Benavente	30.484	30.901	417
C/Fernando el Católico nº19	41.227	41.640	413
C/Jaime el Conquistador nº30	50.194	50.603	409
Avenida de Alfonso XII nº 54	21.408	21.753	345
Paseo de los Olmos nº28	31.976	32.302	326

Los datos muestran que la zona en torno al río Manzanares agrupa a las estaciones con mayor número de asimetría positiva. Otras zonas que aparecen son la Avenida de Alfonso XII nº 54, que sorprende debido a lo alejada que está y el centro histórico con estaciones como la Plaza de Jacinto Benavente.

Tabla 7. Estaciones con mayor número de asimetría negativa total contabilizando empleados

Nombre de la estación	**Número de salidas efectuadas/desenganches**	**Número de llegadas efectuadas/enganches**	**Asimetría total**
C/Velázquez nº 130	38.840	38.029	-811
C/ del General Yagüe nº 57	26.722	26.044	-678
C/José Ortega y Gasset nº 30	32.224	31.623	-601
Paseo del Prado nº 1	24.871	24.278	-593
Calle Serrano nº 34 - A	10.961	10.382	-579
Paseo de la Castellana nº 67	18.119	17.577	-542
Paseo de la Castellana nº 4	13.157	12.646	-511
C/Velázquez nº 75	18.376	17.876	-500
C/Serrano nº 34 - B	12.501	12.087	-414
C/Goya nº 18	25.666	25.264	-402

Los datos nos indican que la zona céntrica del Paseo de la Castellana y las calles más cercanas a la misma son las zonas más perjudicadas y con menos disponibilidad de bicicletas después de tener en cuenta el rebalanceo por parte de los empleados. En específico, y lejos de este grupo de estaciones sorprende la aparición de C/ del General Yagüe nº57, esto es lógico debido a ser la estación más alejada de la zona de Tetuán. Este fenómeno se puede ver en la ilustración 24, donde las estaciones de alrededor también cuenta con una asimetría negativa en torno a -200.

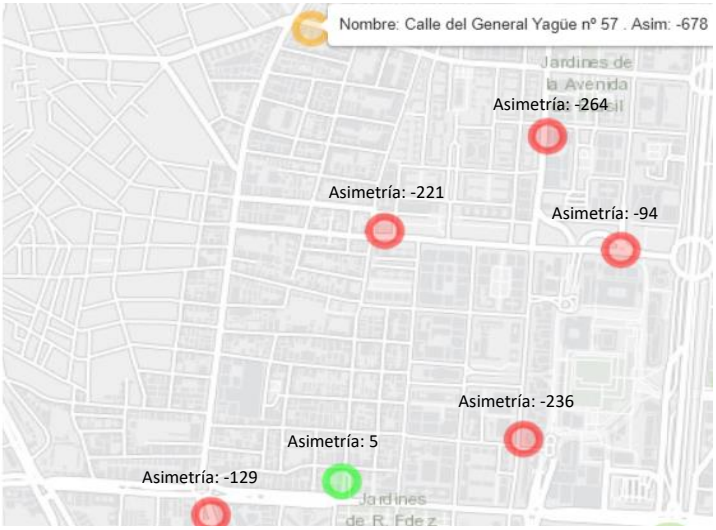


Ilustración 24. BiciMAD – Mapa de asimetrías de la estación General Yagüe nº57 y alrededores (Elaboración propia).

4.5.4 Mapas de asimetría total y densidad

El mapa de densidad, ilustración 25, se ha elaborado con una muestra del 25% debido a la capacidad de procesamiento computacional disponible. El mapa de densidad comprende la actividad de las zonas donde los usuarios han realizado el desenganche de la bicicleta e iniciado su recorrido, con mayor o menor frecuencia. Las zonas más claras nos indican una mayor actividad y en este caso corresponden a la zona del Centro, el barrio de Salamanca, Chamberí centro y sur además de la estación de la calle Cea Bermúdez nº59, de la calle Santa Engracia nº127 (por su proximidad a Ciudad Universitaria) y de la calle Velázquez nº130. Se ha decidido realizar el mismo mapa de densidad por número de enganches, pero el resultado ha sido muy similar. Esto se debe a que las estaciones con mayor número de salidas y entradas suelen ser las zonas del Centro, produciendo que las zonas más alejadas de estas tengan menor actividad.

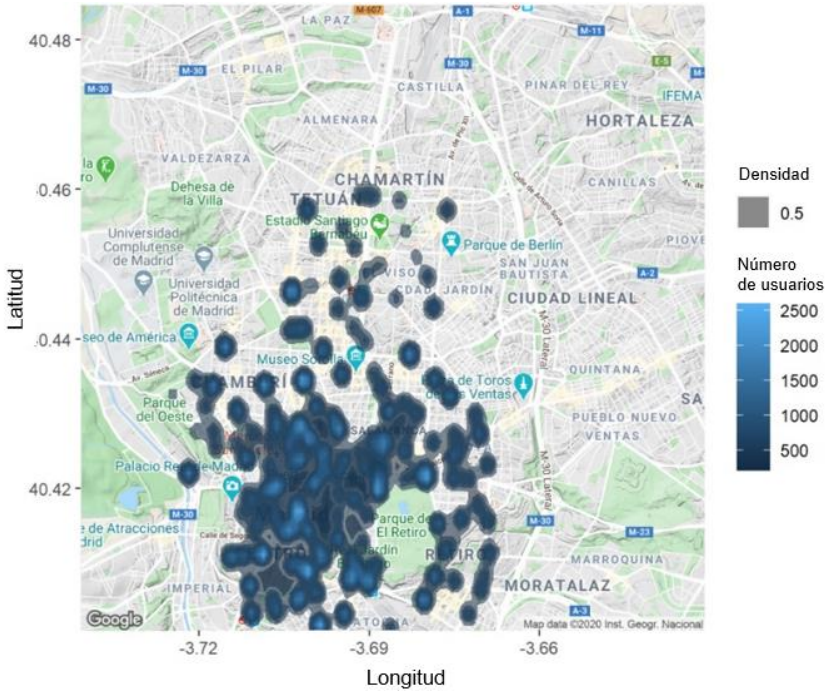


Ilustración 25. BiciMAD – Mapa de densidad por número de desenganches de todas las estaciones (Elaboración propia).

En el mapa de asimetría total, ilustración 26, se ha clasificado el número de asimetría con diversos colores. Debido a la cantidad de estaciones se ha elegido realizar grupos de estas por cercanía geográfica que se muestran en amarillo (si en esa zona hay más de 10 estaciones) y en verde (si en esa zona hay menos de 10 estaciones) en el mapa. Se decide realizar otro mapa con las veinte estaciones con mayor (10 estaciones) y menor (10 estaciones) número de asimetría con el fin de ver mejor los patrones que nos ofrecen estos mapas.

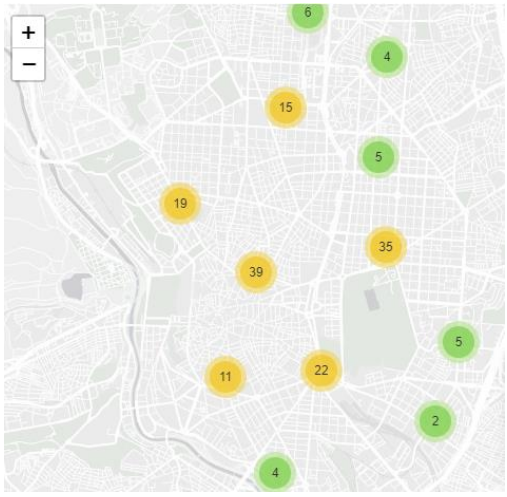


Ilustración 26. BiciMAD – Mapa de asimetría total, clasificación por número de asimetría y clústeres por cercanía geográfica (Elaboración propia).

En este último mapa de asimetría, ilustración 27, se localizan las diez estaciones con mayor y menor asimetría. Se visualiza un patrón muy claro, con la zona del barrio de Salamanca con una asimetría muy negativa (más salidas que entradas) y la zona del centro y Príncipe Pío con asimetrías muy positivas (más entradas que salidas). También destaca la estación de la calle del General Yagüe nº 57 en Tetuán con una asimetría negativa y muy alejada, como ya habíamos detectado previamente. En este mapa, también se ha realizado una agrupación por cercanía geografía en amarillo (si en esa zona hay más de 10 estaciones) y en verde (si en esa zona hay menos de 10 estaciones).

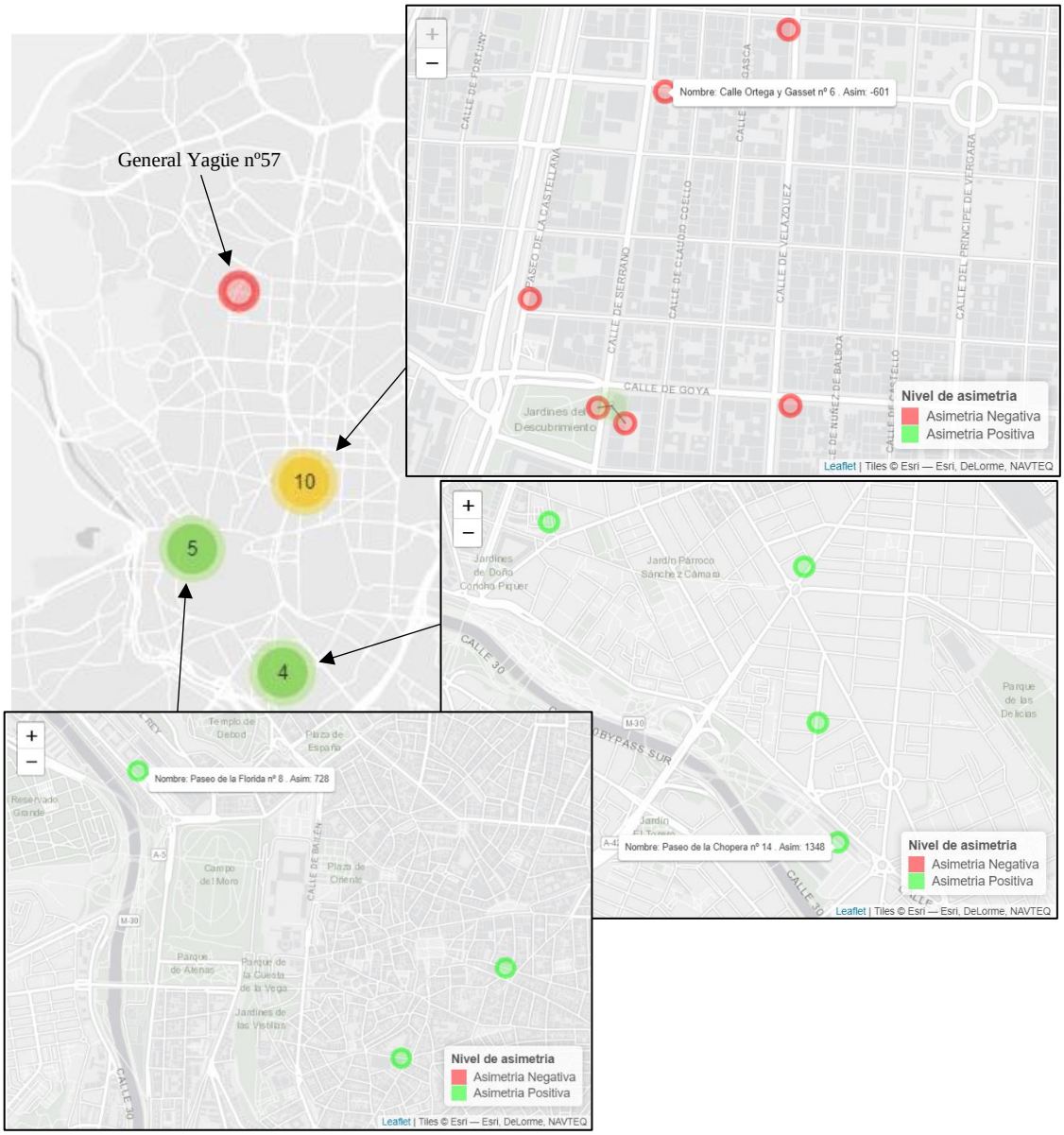


Ilustración 27. BiciMAD – Mapa de asimetría de las 20 estaciones con mayor y menor asimetría y clústeres por cercanía geográfica (Elaboración propia).

Como conclusión a los mapas realizados y el análisis de asimetría, en 2018, la zona de mayor actividad es la del Centro. Las asimetrías positivas o las estaciones destino son aquellas localizadas en el sur de Madrid, cerca de Madrid Río. También son estas estaciones las más utilizadas con mayor número de salidas y de entradas que aquellas en la Castellana, donde se da una fuerte asimetría negativa. Esto da lugar a un problema en la zona de la Castellana, Velázquez, Serrano etc. ya que no son las estaciones más utilizadas, sufren mayores salidas que entradas y están alejadas del Centro, lo que hace más costoso el rebalanceo para los empleados.

4.5.5 Rutas y estaciones más populares por usuarios y empleados

En este apartado se explorará cuáles son las estaciones de salida y entrada con mayor número de viajes para los usuarios abonados y ocasionales. También se exploran las rutas más realizadas por los empleados para paliar el desbalanceo del sistema.

Tabla 8. Estaciones de origen más populares por número de viajes efectuados por abonados y ocasionales en 2018

Estación origen del viaje por los usuarios	**Número de salidas anuales efectuadas**
C/Valencia nº1	47.055
C/del General Álvarez de Castro nº2	46.480
C/Jaime el Conquistador nº30	44.320
Plaza de la Cebada nº16	44.091
C/Santa Engracia nº168	41.723
Avenida de Menéndez Pelayo nº11	40.612
Paseo de Santa María de la Cabeza nº58	39.008
Paseo de la Chopera nº14	38.290
C/Cea Bermúdez nº59	36.761
C/Fernando el Católico nº19	36.611
C/Barceló nº7	36.481
Plaza de la Independencia nº6	34.486
Plaza de Alonso Martínez nº5	33.528
Plaza de Pedro Zerolo	33.181
Avenida de Menéndez Pelayo nº38	33.027
Puerta del Sol nº1 - A	32.317
Glorieta de la Puerta de Toledo nº1	32.052
C/Velázquez nº130	31.600

C/Juan Bravo nº50	31.018
Plaza de San Ildefonso nº3	30.995

C/Valencia nº1 (Lavapiés), C/del General Álvarez de Castro nº2 (Tetuán), C/Jaime el Conquistador nº30 (Centro sur), Plaza de la Cebada nº16 (La Latina) y C/Santa Engracia nº168 (Cerca de Ciudad Universitaria), son aquellas estaciones con mayor número de salidas en 2018. La mayoría de estas estaciones en esta lista se concentran en las zonas descritas entre paréntesis.

Tabla 9. Estaciones de destino más populares por número de viajes efectuados por abonados y ocasionales en 2018

Estación destino del viaje por los usuarios	**Número de llegadas anuales efectuadas**
Paseo de la Chopera nº14	52.677
Calle Jaime el Conquistador nº30	48.471
Plaza de la Cebada nº16	48.101
Calle del General Álvarez de Castro nº2	47.203
Paseo de Santa María de la Cabeza nº58	45.738
Calle Valencia nº1	45.051
Calle Santa Engracia nº168	41.780
Avenida de Menéndez Pelayo nº11	39.520
Calle Fernando el Católico nº19	39.430
Calle Cea Bermúdez nº59	37.340
Plaza de la Independencia nº6	35.818
Calle Barceló nº7	35.502
Avenida de Menéndez Pelayo nº38	35.486
Paseo de la Florida nº8	35.189
Glorieta de la Puerta de Toledo nº1	33.577
Plaza de Alonso Martínez nº5	33.278
Plaza de Pedro Zerolo	33.119
Calle Velázquez nº130	32.303
Calle Palos de la Frontera nº40	31.548
Plaza de San Ildefonso nº3	31.162

Paseo de la Chopera nº14, Jaime el Conquistador nº30 y Paseo de Santa María de la Cabeza son zonas que pertenecen al Centro sur, a lo largo del río Manzanares. La calle

del General Álvarez de Castro nº2 es también una estación muy popular de destino. Esto indica un fuerte predominio de uso alrededor de todo el centro.

Tabla 10. Estaciones de origen más populares por número de viajes efectuados por empleados en 2018

Estación origen del viaje por los usuarios	**Número de salidas anuales efectuadas**
Paseo de la Chopera nº14	13.452
Paseo de Santa María de la Cabeza nº58	9.636
Paseo de la Florida nº8	9.364
C/Cerro de la Plata nº2	7.643
C/Velázquez nº130	7.116
Paseo de los Olmos nº28	6.101
Plaza de la Cebada nº16	5.910
C/Jaime el Conquistador nº30	5.800
Avenida de Menéndez Pelayo nº38	5.254
Plaza de la Independencia nº6	5.226
Avenida de Menéndez Pelayo nº90	4.695
Calle Fernando el Católico nº19	4.532
Calle Ortega y Gasset nº6	4.409
Paseo de las Delicias nº92-94	4.267
Calle Palos de la Frontera nº40	3.950
Carrera de San Francisco nº1	3.766
Plaza de San Francisco nº5	3.528
Paseo del Prado nº1	3.511
Calle Antonio Maura nº15	3.368
Paseo de la Esperanza nº2	3.182

Las estaciones de origen de viaje más populares también son aquellas en torno al Río Manzanares, lo que sugiere que son rutas muy populares, como veremos a continuación. En esta tabla, están presentes estaciones muy cerca de la zona de Atocha, como es el caso de C/Cerro de la Plata nº2, Avenida de Menéndez Pelayo nº90, Avenida de Menéndez Pelayo nº38 o Paseo de las Delicias nº92-94, posiblemente debido a su cercanía a la estación de Atocha y sus múltiples conexiones de transporte público a todo Madrid.

Tabla 11. Rutas más populares por número de viajes efectuados por abonados y ocasionales en 2018

Estación origen	**Estación destino**	**Número de viajes efectuados entre ambas estaciones**
Plaza de Alonso Martínez nº5	Calle Santa Engracia nº168	2.396
Calle Valencia nº1	Paseo de la Chopera nº14	1.950
Paseo de la Chopera nº14	Paseo de la Florida nº8	1.908
Paseo de la Florida nº8	Paseo de la Chopera nº14	1.766
Glorieta de Embajadores nº2	Paseo de la Chopera nº14	1.707
Calle Santa Engracia nº14	Calle Santa Engracia nº168	1.667
Calle Jaime el Conquistador nº30	Plaza de la Cebada nº16	1.652
Calle Jaime el Conquistador nº30	Paseo de la Chopera nº14	1.646
Calle del General Álvarez de Castro nº2	Calle Santa Engracia nº168	1.616
Paseo de la Chopera nº14	Calle Jaime el Conquistador nº30	1.578
Calle del General Álvarez de Castro nº2	Calle Cea Bermúdez nº59	1.565
Calle Cea Bermúdez nº59→	Calle del General Álvarez de Castro nº2	1.503
Glorieta de la Puerta de Toledo nº1	Paseo de la Chopera nº14	1.503
Plaza de la Cebada nº16	Paseo de la Chopera nº14	1.493

Las rutas más populares, tanto por los abonados anuales como por los usuarios ocasionales son rutas en torno al río Manzanares. La ruta más popular, sin embargo, es desde la plaza de Alonso Martínez nº5 hasta la calle Santa Engracia nº168. Esto puede ser debido a que ambas estaciones se encuentran en línea recta y a que Alonso Martínez se encuentre en una zona limítrofe al barrio de Salamanca y la zona Centro. La estación que más destaca en todos los aspectos es Paseo de la Chopera nº14, cerca de la entrada al centro de exposiciones Matadero Madrid en Madrid Río.

Tabla 12. Rutas más populares por número de viajes efectuados por empleados en 2018

Estación origen	**Estación destino**	**Número de viajes efectuados entre ambas estaciones**
Paseo de la Chopera nº14	Calle Velázquez nº130	555
Paseo de Santa María de la Cabeza nº58	Calle Velázquez nº130	484
Paseo de la Chopera nº14	Paseo del Prado nº1	412
Paseo de la Chopera nº14	Paseo de la Castellana nº42	404
Paseo de la Chopera nº14	Paseo de la Castellana nº4	398
Paseo de Santa María de la Cabeza nº58	Calle Goya nº18	394
Paseo de la Chopera nº14	Paseo de Recoletos nº32-34	365
Calle Cerro de la Plata nº2	Calle Velázquez nº130	349
Paseo de la Chopera nº14	Calle Goya nº18	344
Paseo de la Chopera nº14	Calle Velázquez nº75	341
Paseo de la Chopera nº14	Plaza de la Independencia nº6	341
Paseo de Santa María de la Cabeza nº58	Calle Serrano nº34 - A	340
Paseo de la Florida nº8	Calle Goya nº18	339
Paseo de Santa María de la Cabeza nº58	Paseo de Recoletos nº32-34	338
Calle del General Yagüe nº57	Calle Orense nº36	334
Plaza de Independencia nº6	Avenida de Menéndez Pelayo nº90	331
Calle Santa Engracia nº168	Paseo de la Castellana nº67	325
Paseo de la Chopera nº14	Calle Ortega y Gasset nº6	323
Calle Cerro de la Plata nº2	Calle Goya nº18	321
Paseo de Santa María de la Cabeza nº58	Paseo de la Castellana nº4	320

Las rutas para llevar a cabo el rebalanceo el sistema son las siguientes. Si observamos con detenimiento, coincide con el análisis de asimetría llevado a cabo, donde los empleados recogen las bicicletas de estaciones como Paseo de la Chopera nº14, de gran asimetría positiva, con mucha frecuencia y las distribuyen hacia la zona de la Castellana y el barrio de Salamanca, de gran asimetría negativa. También podemos darnos cuenta en la tabla anterior con la estación Paseo de la Chopera nº14 como estación destino con mayor frecuencia, lo que sugiere que esta estación es clave para el rebalanceo del sistema.

4.5.6 Ejemplo: Cálculo matemático para ver la eficiencia de los empleados mediante la disminución o aumento de asimetría de una estación

De una manera muy práctica y simple seleccionaremos una estación**,** Paseo de la Chopera nº14, que tiene una asimetría extremadamente positiva como ya hemos visto en el anterior análisis estacional en el apartado 4.5.1, para ver desde una perspectiva matemática cómo los empleados actúan y realizan el rebalanceo de esta estación utilizando los datos analizados. Este análisis es meramente una estimación, y puede variar, dependiendo de cómo se filtren los datos. Una asimetría muy positiva significa que hay un exceso en el número de usuarios que dejan la bicicleta aquí, descompensando el sistema en otras estaciones que reciben menos bicicletas. Por lo que los empleados deberán coger las bicicletas (reduciendo así la asimetría) y distribuirlas a otras estaciones.

La fórmula para calcular la asimetría total que tendría la estación de no haber empleados sería la suma de los números de asimetría de las cuatro estaciones del año:

Número de Asimetría de estación X SIN empleados = Σ Núm. Asimetría de estación X en todas las estaciones del año = Núm.Asim. Invierno + Núm.Asim.Primavera + Núm.Asim.Verano + Núm.Asim.Otoño

Por lo que,

Número de Asimetría de Paseo de la Chopera nº14 SIN empleados = 2.579+3.702+4.610+4.498 = 15.389

De la misma manera, en el apartado 4.5.3, se estima la asimetría de la estación con la labor de los empleados:

Núm. de Asimetría de estación X CON empleados = Nº llegadas de estación X Total - Nº salidas de estación X Total

Núm. de Asimetría de Paseo de la Chopera nº14 CON empleados =

53.274-51.926 = 1.348

Por lo que los empleados reducen la asimetría para esta estación en:

Núm. de Asimetría de Paseo de la Chopera nº14 SIN empleados - Núm. de Asimetría de Paseo de la Chopera nº14 CON empleados = 15.389 – 1.348 = 14.041 viajes desde Paseo de la Chopera donde las bicicletas son distribuidas a otras estaciones en 2018.

La distribución de bicicletas diaria se puede calcular de la siguiente manera:

(*Núm. de Asimetría de Paseo de la Chopera nº14 SIN empleados - Núm. de Asimetría de Paseo de la Chopera nº14 CON empleados) / Días operativo donde asumimos la base 365 días*

14.041 bicicletas /365 días = 38,47 bicicletas /día se distribuyen desde Paseo de la Chopera nº14 a otras estaciones del sistema.

De la misma forma, también podemos calcular el reajuste de los empleados en forma porcentual:

(*Núm. de Asimetría de Paseo de la Chopera nº14 SIN empleados - Núm. de Asimetría de Paseo de la Chopera nº14 CON empleados*) / *Núm. de Asimetría de Paseo de la Chopera nº14 SIN empleados*

(15.389 – 1.348) /15.389 = 0,9124 o una disminución del 91,24% en la asimetría de Paseo de la Chopera nº14

Por último, en el apartado 4.5.5 se conocen las salidas realizadas por los empleados para Paseo de la Chopera nº14, lo que sugiere un número menor de salidas, 13.452 en vez de las 14.041 calculadas previamente. Las diferencias se deben a la transformación de datos y la eliminación de falsos viajes, como, por ejemplo, que una bicicleta se desenganche y enganche en menos de 5 segundos en la misma estación. Se ha establecido un filtro donde se admite la misma estación de origen y destino en caso de que la duración del viaje sea superior al minuto y medio o los 90 segundos.

13.452/15.389 =0,8741 o una disminución del 87,41% en la asimetría de Paseo de la Chopera nº14 tras la eliminación de algunos errores

Esto sugiere un fantástico trabajo realizado en esta estación en 2018, entre un 87,41% a un 91,24% en la reducción de asimetría, pero habría que realizar este mismo trabajo para las 172 estaciones restantes, con el fin de evaluar el desempeño agregado de los empleados.

4.6 Análisis del estado de estaciones para el mes de septiembre de 2018 con modelos clustering no supervisados.

¿Se comportan las estaciones de manera diferente, respecto a su capacidad, ocupación y número de reservas? ¿Cómo podemos realizar un mejor balanceo del sistema, de tal forma que se pueda mejorar la gestión incrementando la eficiencia? Para abordar estas preguntas se aplicarán los algoritmos de "clustering" explicados en el apartado 2.3.1 y 2.3.2 de este trabajo y utilizará la base de datos 2, estado de estaciones para septiembre de 2018.

4.6.1 Formación de clústeres por k-Means, según siete variables.

El objetivo de realizar "clustering" por K-means es el de hallar subgrupos o clústeres según siete variables para cada estación: la media porcentual de ocupación de bicicletas ancladas, la media porcentual de bases libres, la capacidad porcentual media operativa de las bases, el número promedio de bases inoperativas, el número promedio de bases inoperativas y el número de reservas medio. Este "clustering" se realiza para encontrar subgrupos de estaciones que se comporten de manera diferente o similar, con el fin de dar recomendaciones respecto a su tratamiento operativo (respecto a la ocupación, el mantenimiento y el número de reservas medio) para cada grupo de estaciones. Junto a estas variables, también se han incluido la longitud y latitud de cada estación, aunque no se han tenido en cuenta a la hora de realizar los clústeres, pero sí, al realizar el mapa de estos en la ilustración 32.

```
## 'data.frame':    172 obs. of  10 variables:
##  $ nombre.estacion       : Factor w/ 172 levels "Agustin de Betancourt".
##  $ avg.ocupacion.ancladas: num  0.298 0.424 0.462 0.234 0.357 ...
##  $ avg.bases.libres      : num  0.702 0.576 0.536 0.766 0.643 ...
##  $ avg.capacidad.estacion: num  0.981 0.984 0.983 0.97 0.973 ...
##  $ avg.bases.inoperativas: num  0.449 0.381 0.452 0.713 0.659 ...
##  $ avg.bases.operativas  : num  23.6 23.6 26.5 23.3 23.3 ...
##  $ avg.bases.totales     : num  23.9 23.9 24.5 23.8 23.8 ...
##  $ longitud.estacion     : num  -3.7 -3.68 -3.68 -3.67 -3.69 ...
##  $ lat.estacion          : num  40.4 40.5 40.4 40.4 40.4 ...
##  $ avg.numero.reservas   : num  0.1057 0.0751 0.057 0.0223 0.0974..
```

Ilustración 28. Estructura de la base de datos utilizada en R para la generación de clústeres

Se ha transformado la base de datos inicial mediante tablas dinámicas en Excel (Elaboración propia).

A continuación, se define cómo se han calculado los registros para cada estación y cómo se han creado las variables, con formato de media, en la base de datos. Para más detalle sobre cómo se han realizado las transformaciones de datos diríjase al anexo II.

Se definen bases inoperativas y operativas como;

Bases inoperativas por estación, día y hora =

bases.totales (total_bases) – (ocupación.ancladas (dock_bikes) + bases.libres (free_bases))

Bases operativas por estación, día y hora =

bases.totales (total_bases) – Bases inoperativas por estación, día y hora

Por lo que un registro de ocupación, base libre o capacidad se computa como;

Ocupación bicicletas (%) por estación, día y hora =

ocupación.ancladas (dock_bikes)/Bases operativas por estación, día y hora

Bases libres (%) por estación, día y hora =

ocupación.ancladas (dock_bikes)/Bases operativas por estación, día y hora

Capacidad de la estación (%) por estación, día y hora =

Bases operativas por estación, día y hora / bases.totales (total_bases)

Teniendo en cuenta las anteriores fórmulas, para cada estación hay aproximadamente 720 registros o mediciones (30 días x 24 horas/día), por lo que se pueden computar las medias mensuales por estación como:

Media del número de bases inoperativas de estación X (avg.bases.inoperativas) =

$$\frac{1}{719}\sum_{i=1}^{719} Bases\ inoperativas\ por\ día\ y\ hora_i$$

Media del número de bases inoperativas de estación X (avg.bases.operativas) =

$$\frac{1}{719}\sum_{i=1}^{719} Bases\ operativas\ por\ día\ y\ hora_i$$

Media del número de reservas de estación X (avg.bases.operativas) =

$$\frac{1}{719}\sum_{i=1}^{719} Número\ de\ reservas\ por\ día\ y\ hora_i$$

Ocupación de las bicicletas media (%) de estación X (avg.ocupación.ancladas) =

$$= \frac{1}{719}\sum_{i=1}^{719} Ocu.\ bicicletas\ (\%)\ por\ día\ y\ hora_i$$

Media de bases libres sin bicicletas (%) de estación X (avg.bases.libres) =

$$= \frac{1}{719}\sum_{i=1}^{719} Bases\ libres\ (\%)\ por\ día\ y\ hora_i$$

Capacidad media (%) de estación X (avg.capacidad.estación) =

$$= \frac{1}{719}\sum_{i=1}^{719} Bases\ libres\ (\%)\ por\ día\ y\ hora_i$$

Una vez definida la base de datos a utilizar, aplicamos el método de Elbow y Silhouette, ya que utilizando K-means, debemos predefinir el número de clústeres a realizar. Estos métodos ya han sido explicados en el apartado 2.3.3 de este trabajo. Los dos métodos indican 5 clústeres o grupos cómo el número óptimo, por lo que no tenemos la necesidad de utilizar ningún otro indicador.

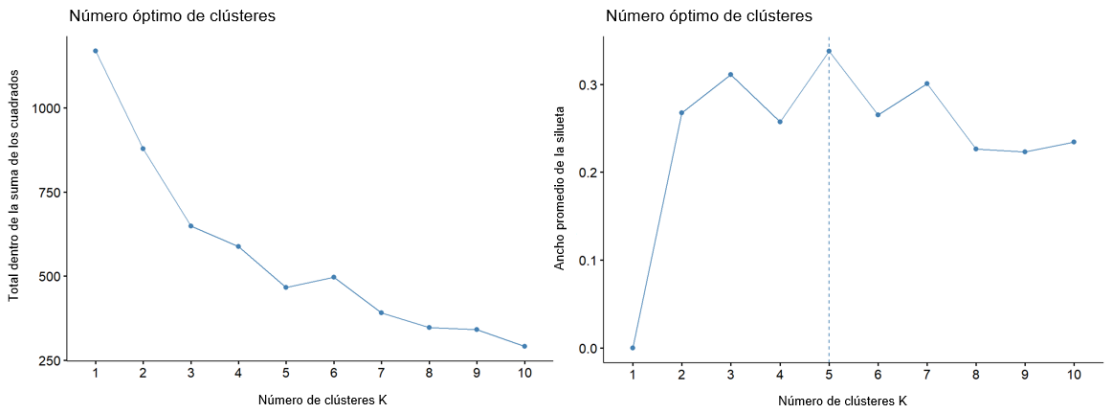


Ilustración 29. BiciMAD – Métodos Elbow y Silhouette para elegir el número de clústeres (Elaboración propia).

Tras decidir el número de clústeres a formar, se deciden calcular las distancias entre las estaciones mediante la distancia euclidiana. El resultado es una matriz de distancias que indica cómo de parecidos o diferentes son cada par de observaciones/estaciones. Esta matriz de distancia, ilustración 30, muestra algunas áreas rojas donde las estaciones son muy parecidas entre sí, lo que significa una tendencia a agruparse y, por lo tanto, significa que K-means es un método correcto para evaluar y formar grupos en el conjunto de datos.

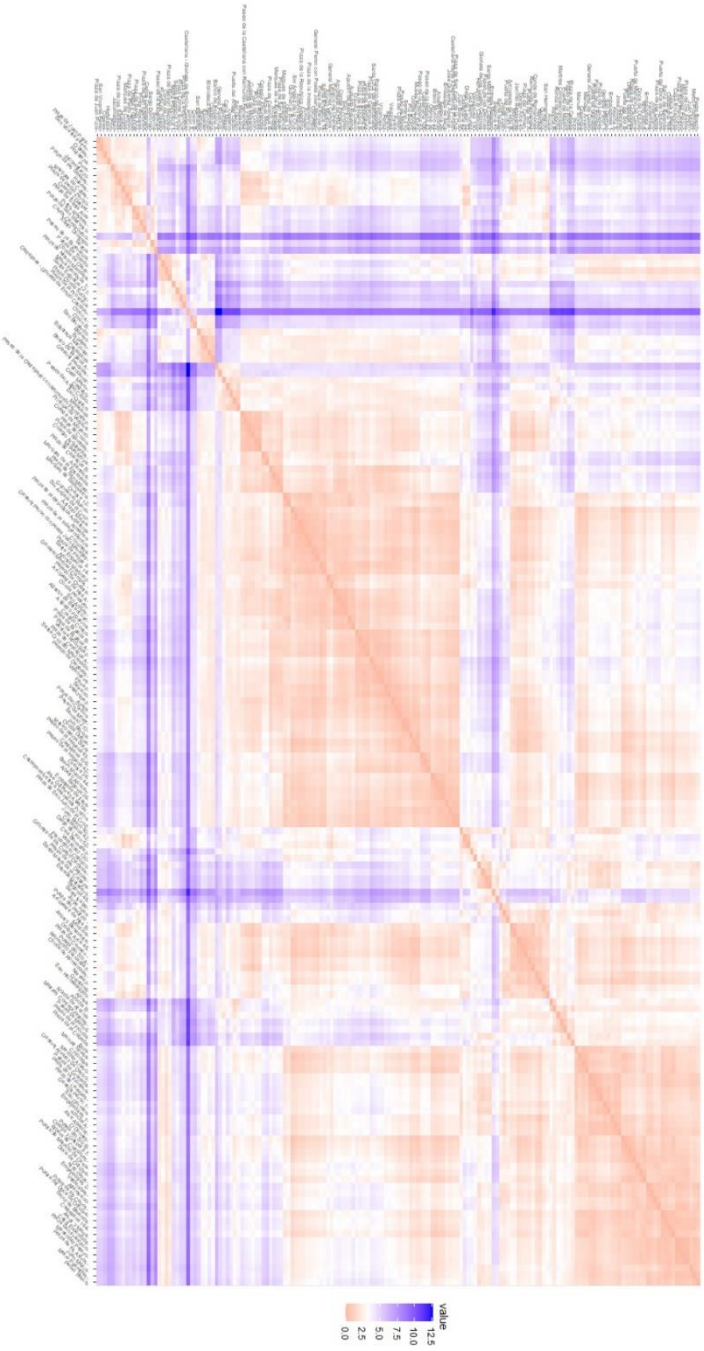


Ilustración 30. BiciMAD – Matriz de distancias de "clustering" K-means (Elaboración propia).

Si ahora aplicamos el algoritmo K-means sobre las 7 variables anteriormente mencionadas, se podrán observar los 5 grupos resultantes en sus dos primeras componentes principales. El Análisis de Componentes Principales (PCA) realiza una proyección con el fin de que los datos queden mejor representados en términos de mínimos cuadrados, reduciendo la dimensionalidad de los datos. Por ejemplo, la primera dimensión PCA en la ilustración 31, es una variable sintética formada a partir de una combinación lineal de las 7 variables que se han utilizado para realizar el "clustering" por K-means. Las dos primeras componentes principales o dimensiones de PCA agrupan la mayoría de la varianza (65,5% en este caso). El Análisis de Componentes Principales queda fuera del alcance de este TFG, pero es útil para ver un resumen de los clústeres formados por K-means.

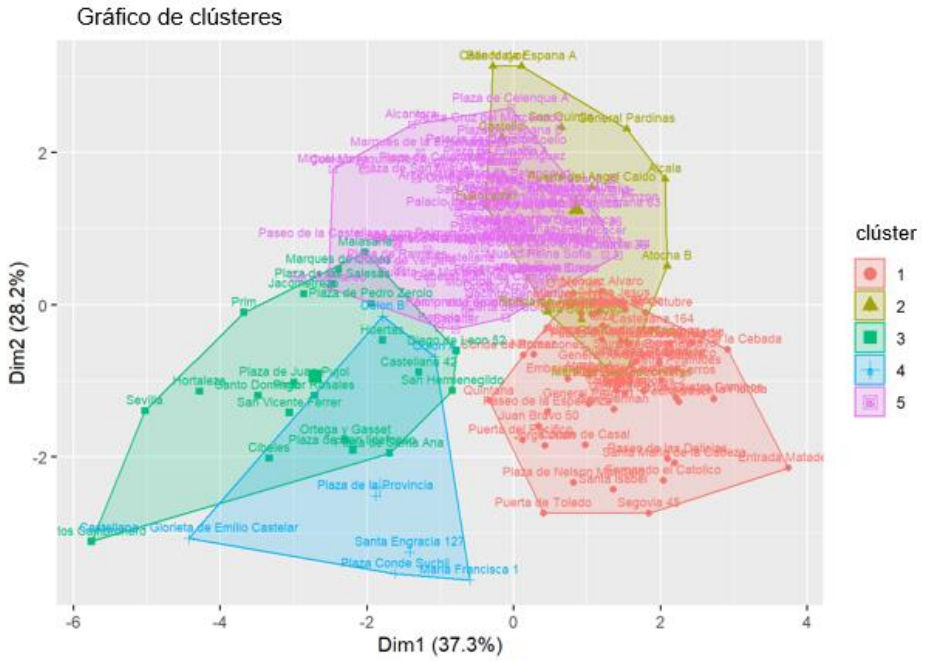


Ilustración 31. BiciMAD – Gráfico de "clustering" por K-means en las dos primeras componentes principales (Elaboración propia).

Los centroides de los diferentes grupos del nuestro modelo K-means están descritos en la siguiente tabla:

Tabla 13. Centroides del modelo estático mediante K-means.

Grupo	avg.ocu.ancladas	avg.bases.lib	avg.capacidad.estac	avg.bases.inop.	avg.bases.op	avg.bases.tot	avg.num.rva
1	53,39%	46,58%	96,51%	0,83	23,09	23,82	0,12
2	40,54%	59,44%	95,69%	1,18	26,20	24,63	0,09
3	32,16%	67,09%	90,49%	2,23	21,20	23,70	0,09
4	44,00%	55,86%	95,28%	0,83	16,60	22,30	0,07
5	34,87%	65,10%	96,99%	0,72	23,08	23,79	0,06

En verde, el mejor registro dentro de los 5 grupos. En rojo, el peor.

Podemos concluir que:

- El grupo 1 está formado por aquellas estaciones con una ocupación media por estación del 53% (asimetría positiva, ya que llegan más bicicletas de las que salen) y donde la capacidad por estación es mayor, lo que sugiere un mejor mantenimiento que los demás grupos. También destaca por un número mayor de reservas.

- El grupo 2 destaca por un número mayor en las bases operativas y totales, estas estaciones conforman zonas céntricas donde se necesitan mayor número de bases/enganches.

- El grupo 3 tiene los peores registros. En ocupación media por estación, el grupo 3 tiene tan solo un 32% (asimetría negativa) y en capacidad disponible, estas estaciones solo obtienen un 91% lo que sugiere un alto número de bases inoperativas, como es el caso, con 2,2 bases por estación de media. Se debe realizar mayor mantenimiento a estas estaciones.

- El grupo 4 son aquellas zonas con menos bases totales de media, lo que sugiere zonas más residenciales y de menos tránsito. Destaca una elevada ocupación media, por lo que también podrían servir para realizar el rebalanceo del sistema.

- Por último, el grupo 5 comprende aquellas estaciones con un mantenimiento superior, con poca ocupación, un 35%, una gran capacidad disponible y el número de bases inoperativas y reservas más bajo.

Respecto al número de estaciones en cada clúster o grupo, el grupo 1 está formado de 55 estaciones, el grupo 2 por 13 estaciones, el grupo 3 por 21 estaciones, el grupo 4 por únicamente 7 estaciones y el grupo 5 agrupa la mayoría, con 72 estaciones. Tanto los métodos Elbow y Silhouette como el algoritmo K-means nos explican que las estaciones presentan cinco comportamientos diferentes respecto a las variables utilizadas.

Se ha notado algo de solapamiento entre los clústeres formados. Esto se debe a que no hay tanta diferencia entre algunas observaciones de cada grupo, por lo que en un futuro análisis, se podría elegir un número menor de clústeres o realizar medias semanales en vez de mensuales para conocer con más detalle el comportamiento de cada estación.

A continuación, en la ilustración 32 se enseña el mapa realizado con la API de Google Maps en Google Cloud, de los clústeres por K-means. Curiosamente, aquellas estaciones con mayor ocupación media de bicicletas (Clúster 1) son las mismas que presentan una asimetría muy positiva, cuando se utilizaba la base de datos 1, en la ilustración 27 de este trabajo.

Ilustración 32. BiciMAD – Gráfico de la localización geográfica de los clústeres por K-means* (Elaboración propia).
*Los colores del mapa y de la ilustración 31 no coinciden en color, pero sí en número.

En este mapa se puede ver, de forma más clara, que las estaciones del sur deben ser utilizadas por los empleados para reabastecer de bicicletas a las del Centro. Se puede ir más allá y concluir que las estaciones con menor ocupación de bicicletas (menos bicicletas disponibles para los usuarios) deberán ser compensadas con aquellas estaciones con mayor ocupación de bicicletas MÁS CERCANAS (más bicicletas disponibles para los usuarios), siempre teniendo en cuenta el uso de ambas estaciones (no debemos compensar dos estaciones con bicicletas donde una estación es muy utilizada y la otra no lo es). También podemos ver que en la zona de Castellana y Tetúan hay muy pocas estaciones del clúster 1 para compensar al resto, lo que crea un problema a la hora de encontrar bicicletas disponibles para compensar estas estaciones. Por último, de este mapa se puede concluir que las estaciones periféricas son aquellas que registran mayores entradas, mantienen una asimetría positiva y se utilizan para balancear las estaciones del centro, donde se registran mayor número de salidas.

Precisamente para obtener mayor detalle, y para conocer cómo balancear mejor el sistema haciendo referencia al previo análisis de asimetría, se decide realizar un "clustering" de medias de series temporales de 24 horas por estación en el siguiente apartado, descomponiendo esta variable con media mensual total (avg.ocupacion.ancladas), en media mensual por cada hora, ya que utilizando medias mensuales para cada variable, no se perciben realmente todas las diferencias entre las estaciones.

4.6.2 Formación de clústeres de series temporales por "clustering" jerárquico, según ocupación por día y hora de septiembre de 2018

Finalmente, se ha decidido estudiar el comportamiento de las estaciones mediante la variable ocupación realizando una media por hora en vez de una mensual. El objetivo es balancear mejor el sistema por dos criterios; encontrando estaciones que en ocupación se comporten de manera contraria y que estén cerca entre sí.

El siguiente gráfico, ilustración 33, muestra por cada línea el comportamiento medio de cada estación en septiembre de 2018. Se pueden visualizar dos patrones muy visibles. La línea roja muestra estaciones utilizadas por la mañana y tarde. La línea naranja muestra estaciones muy ocupadas al mediodía.

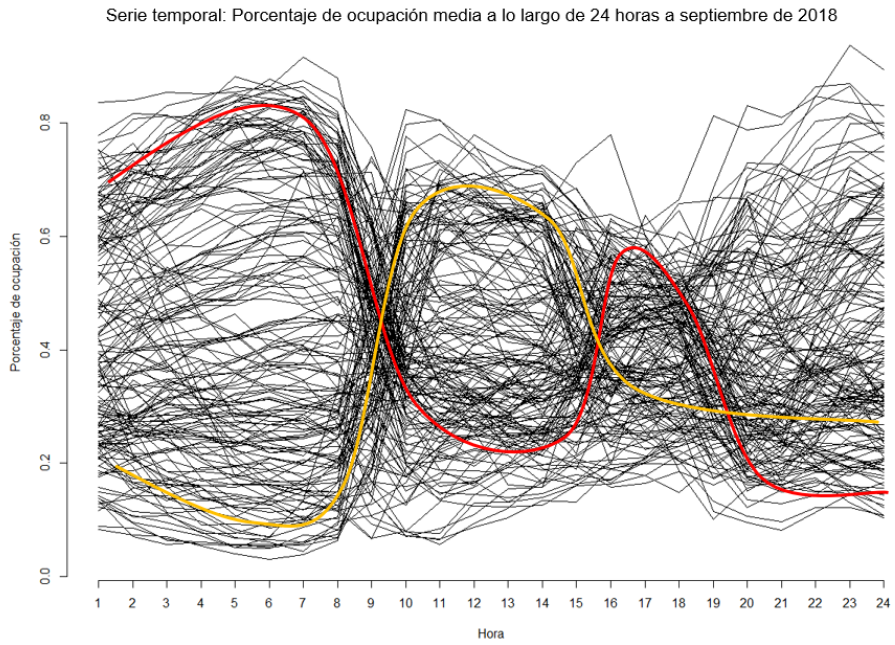


Ilustración 33. BiciMAD – Series temporales de cada estación sobre el porcentaje de ocupación medio para cada hora en septiembre de 2018. Las líneas rojas y naranja han sido realizadas a mano (Elaboración propia).

La idea es realizar "clustering" jerárquico mediante la distancia DTW (Dynamic Time Warping) cuya elección es explicada en el apartado 2.3.2, para formar grupos con las series temporales que más se parezcan por forma entre sí. Una vez realizados los clústeres, se sabe cómo balancear las estaciones, compensando la cantidad de bicicletas por dos criterios, el de cómo de diferentes sean las formas entre las estaciones, para que el sistema se compense de mejor manera y el de cercanía entre estaciones, debido a la eficiencia resultante en tiempo, con el objetivo de que todas las estaciones tengan un número más equitativo de bicicletas.

Se ha decidido agrupar en medias de ocupación por hora (24 registros por estación). Esta decisión viene dada del estudio de tráfico y los sensores de la Agencia de Transporte Sueca anteriormente comentado en el apartado 3.2.1, donde los investigadores agruparon 6 meses de datos en medias horarias para 45 ciudades, para posteriormente realizar "clustering" de series temporales mediante la distancia DTW e identificar las ciudades con patrones de tráfico parecidos. (Verendel & Yeh, 2019)

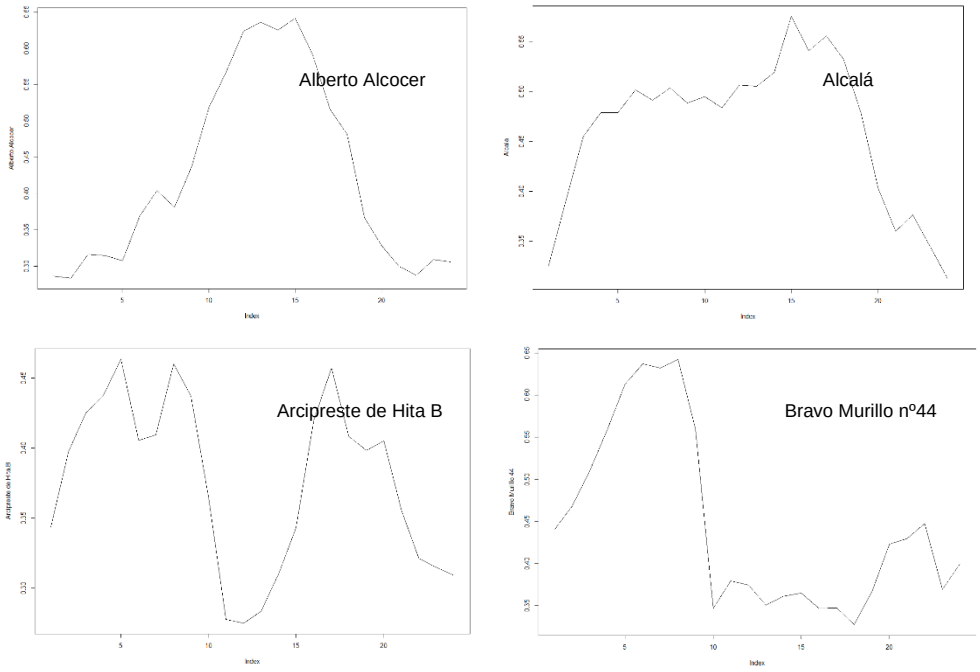


Ilustración 34. BiciMAD – Serie temporales sobre el porcentaje de ocupación medio de bicicletas para cada hora de cuatro estaciones: Alberto Alcocer, Alcalá, Arcipreste de Hita B y Bravo Murillo (Elaboración propia).

De la misma manera, se puede ver el comportamiento individual de las estaciones. En la ilustración 34, se han seleccionado cuatro estaciones de entre las primeras, donde se puede analizar el distinto comportamiento en ocupación para la estación de Alberto Alcocer, Alcalá, Arcipreste de Hita B y Bravo Murillo. Si observamos con detenimiento, las estaciones de Alberto Alcocer y Alcalá exhiben comportamientos muy similares entre sí, pero muy diferentes a Arcipreste de Hita B y Bravo Murillo (que también exhiben un comportamiento bastante similar), lo que nos indica tendencia a agruparse y a formar clústeres diferentes. Como se ha podido comprobar anteriormente, hay dos tipos de patrones en la ilustración 33, lo que sugiere dos clústeres principalmente.

Finalmente, se realiza el dendograma. Se plantea cómo balancear el sistema:

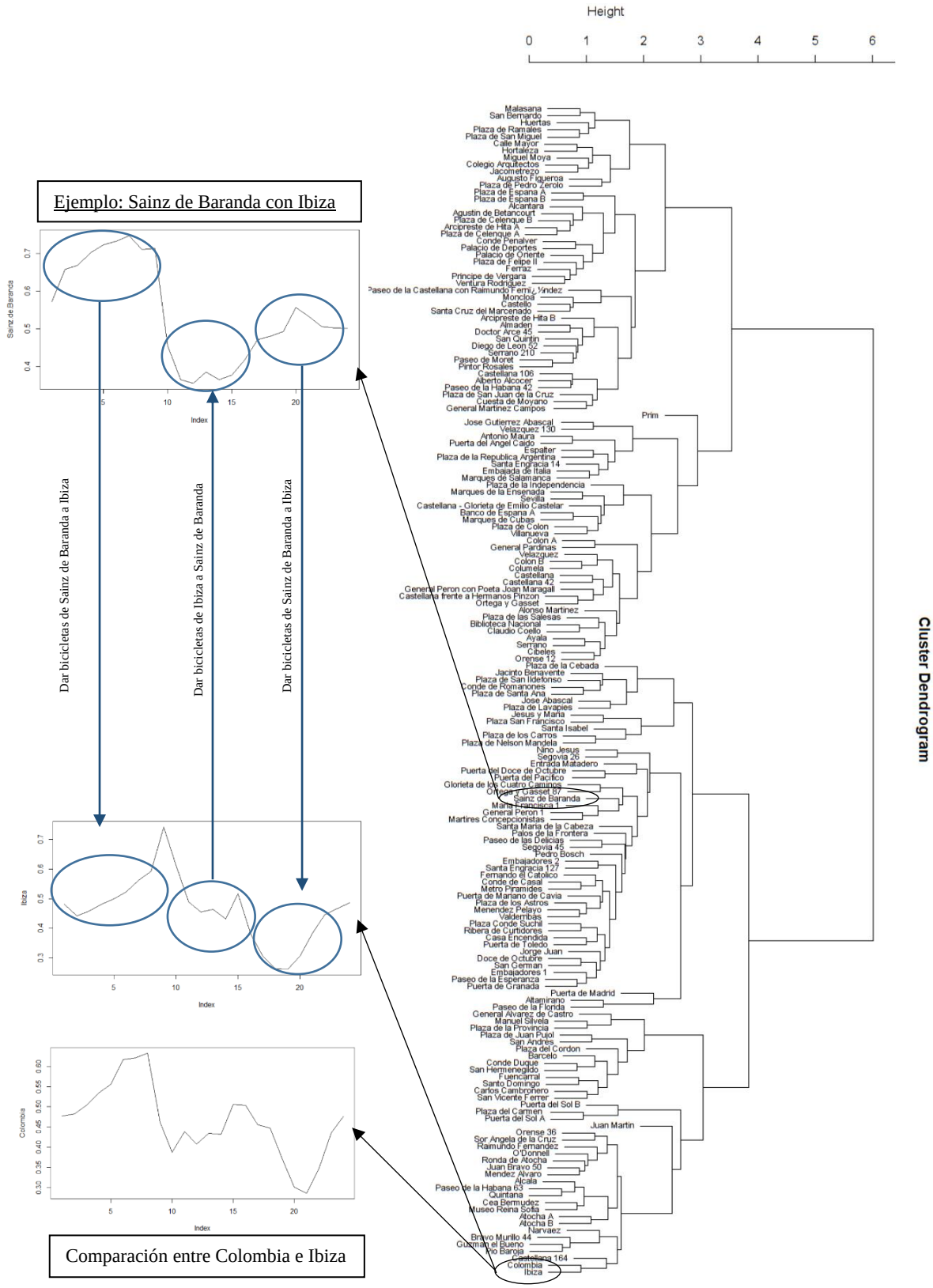


Ilustración 35. BiciMAD – Dendograma de series temporales y ejemplo de rebalanceo (Elaboración propia).

Se perciben dos clústeres principales en el dendograma. El modelo consigue agrupar en distancia DTW y criterio de enlace “average” a las estaciones más parecidas entre sí por su forma, como es el caso de la estación de Ibiza y la de Colombia. En la figura anterior también se ha determinado balancear la estación de Ibiza con la estación de Sainz de Baranda ya que tienen ocupaciones contrarias (cuando una estación tiene una gran ocupación o está muy llena, la otra estará más vacía, por lo que ambas estaciones se pueden dar y recibir bicicletas de manera recíproca para que estén a niveles equitativos). Se ha decidido agrupar estas dos estaciones debido a su cercanía geográfica y a sus diferencias. Ibiza es junto a Pío Baroja, la estación de la zona con asimetría negativa. Sainz de Baranda es una estación muy cercana a Ibiza y tiene una asimetría positiva cómo otras de la misma zona.

El objetivo de esta propuesta es el reparto más equitativo y eficiente posible de las bicicletas, para que ninguna estación se quedé sin ellas, resultando en quejas de los usuarios. No obstante, se ha de mencionar que no es malo que las estaciones se comporten de maneras diferentes en su uso a lo largo de las 24 horas, siempre y cuando se tengan niveles aceptables de bicicletas en todas las estaciones, lo que se intenta lograr con este último “clustering” de series temporales.

4.7 Principales conclusiones sobre el caso BiciMAD

Mediante los múltiples análisis realizados, podemos realizar las siguientes conclusiones:

- Respecto a la demografía del usuario de BiciMAD, se trata de un abonado anual, perteneciente a la franja entre los 27 y 40 años. Se realizan trayectos generalmente cortos y el número de trayectos es alto en los días laborables con picos de actividad por la mañana y tarde. BiciMAD es utilizado fundamentalmente para ir al trabajo y volver del mismo.
- Respecto a la asimetría, las estaciones del sur reciben mayor número de entradas que de salidas, en específico Paseo de la Chopera nº14 y las estaciones de alrededor. Preocupa la zona de la Castellana en 2018 con grandes salidas y pocas entradas. Para realizar el rebalanceo de estas estaciones en el norte se utilizan las estaciones localizadas en el sur, lo que sugiere poca eficiencia y un gran tiempo perdido.

- Respecto a la demanda frente a las variables climáticas, se han podido observar comportamientos no lineales en todas las variables, donde la demanda sube constantemente y a partir de una determinada temperatura o racha de viento, la demanda empieza a bajar. Esto es útil para para predecir la demanda y estimar el número de bicicletas expuestas al público para un determinado día u hora, basado en el pronóstico del clima. Se proponen algunos modelos en el anexo I.
- Por último, la realización de clústeres se ha basado en varios estudios anteriormente realizados para la ciudad de Nueva York. Se ha realizado por tanto un análisis de clústeres de estaciones, como sugiere otro estudio "Rebalancing Bike Sharing Systems: A Multi-source Data Smart Optimization" sobre CitiBike, donde se sugieren varias soluciones para las estaciones menos utilizadas basadas en el establecimiento de precios diferentes. (Liu, Sun, Chen, & Xiong, 2016). Al realizar los clústeres por K-means y realizar un mapa con los mismos, se ha determinado que las estaciones de la periferia se comportan de manera distinta a las de la zona Centro y Paseo de la Castellana.

  En cuanto a los clústeres por series temporales, se concluye que se puede balancear el sistema de mejor manera si se agrupan estaciones por su complementariedad, esto es, que se realicen grupos de estaciones con poca y gran ocupación media, de tal forma que siempre se disponga de bicicletas en ambas estaciones. La realización de estos clústeres para predecir el comportamiento del siguiente año supone seguridad, en caso de que falle el sistema a tiempo real.

5. Conclusiones de este TFG

El objetivo principal de este TFG ha sido el demostrar el potencial de las nuevas tecnologías y el análisis de datos para dar respuesta a problemas que surgen en la movilidad urbana, al cual se ha llegado mediante el desarrollo de un marco teórico y de un caso práctico con datos reales del sistema de bicicletas BiciMAD.

Tras la elaboración del trabajo se ha podido concluir que, en la actualidad, en las empresas surge la necesidad de almacenar grandes cantidades de datos estructurados (SQL) y no estructurados (NoSQL), además de recurrir a métodos de computación paralela como Hadoop y Spark, herramientas de visualización de "Business Analytics" como Tableau, y Power BI y lenguajes de programación como Python y R, para su procesado y posterior

análisis. Además, se han identificado la emisión de gases, la congestión del tráfico, la seguridad vial y el desarrollo del vehículo autónomo como los principales retos de movilidad urbana. Como soluciones a los mismos, las grandes ciudades están adoptando medidas de carácter tecnológico, como la implementación de proyectos piloto o la adopción de las plataformas Mobility-as-a-Service (MaaS), las cuales están basadas en datos.

Por último, con la realización del caso práctico, se ha demostrado la utilidad de los datos mediante múltiples análisis. Se ha podido comprobar cómo el análisis exploratorio nos proporciona conocimiento sobre el comportamiento y las características demográficas de los usuarios, o cómo el clima puede afectar a la demanda. La estadística descriptiva y las herramientas de visualización han demostrado ser de gran utilidad a la hora de comprender y analizar la asimetría o desbalanceo de bicicletas que surge entre las estaciones del sistema, así como para identificar aquellas estaciones con mayor o menor actividad. También, se ha comprobado que el análisis de "clustering" mediante K-means puede ayudar a establecer grupos de estaciones, con el fin de dar un trato operacional diferente a las estaciones. Finalmente, el análisis de "clustering" por series temporales nos ayuda a realizar un rebalanceo más equitativo en las estaciones, de manera que ninguna estación se quede sin bicicletas. Por todo esto, queda demostrado que el uso de datos, su procesamiento y las herramientas de "Business Analytics" y aprendizaje automático son de gran utilidad para entender y resolver problemas en el sector de la movilidad urbana.

Cabe aclarar que este trabajo, y en especial el caso práctico, se ha centrado sólo en algunas de las cuestiones o retos que surgen en los sistemas de transporte de bicicletas, y para ello se han utilizado técnicas de análisis exploratorio y de "clustering". Sin embargo, existen otras muchas posibles aplicaciones de la analítica de datos a este caso concreto. Por eso, como futura línea de investigación, se podría trabajar en la previsión de la demanda diaria, teniendo en cuenta datos de climatología o trabajando con series temporales. Se ha realizado un estudio preliminar en el anexo I. También se propone trabajar con reglas de asociación, otro algoritmo de aprendizaje no supervisado, para determinar la probabilidad de que la bicicleta acabe en una estación habiendo partido de otra, con el fin de que los empleados puedan predecir donde hay una acumulación de bicicletas antes de que se den realmente estos viajes. Otro acercamiento sería la creación de un motor de sugerencias o

recomendaciones, basadas en reglas de asociación, para visitar sitios turísticos en Madrid, con el fin de fomentar el turismo e incrementar el número de usuarios ocasionales.

Se concluye que las empresas de movilidad que analicen y utilicen sus datos constantemente serán empresas ganadoras, ya que son capaces de comprender su negocio o actividad con mayor profundidad y ofrecer cambios más rápidos y eficientes, basados en las preferencias del cliente o usuario.

6. Bibliografía

Accenture. (2019). *Unlocking Data For Smarter Public Transportation.* Estocolmo: Accenture. Obtenido de https://www.accenture.com/_acnmedia/PDF-113/Accenture-Telia-Client-Case.pdf

Achával, L. G. (1950). Éxodo Rural. *Revista de Economía y Estadística, 3*(1-2), 3-30. Obtenido de https://revistas.psi.unc.edu.ar/index.php/REyE/article/download/3266/4864

Antoniou, C., Dimitriou, L., & Pereira, F. (2018). *Mobility Patterns, Big Data and Transport Analytics.* Amsterdam: Elsevier.

Audenhove, F.-J. V., Ali, S., Rominger, G., Salem, J., Arsenyeva, Y., Jaiswal, N., & D'Hooghe, L. (2020). *Rethinking on-demand mobility.* Luxemburgo: Arthur D. Little. Obtenido de https://www.adlittle.com/en/rethinking-demand-mobility

Ayuntamiento de Madrid. (2018). *Madrid Central - Información General.* Obtenido de Ayuntamiento de Madrid - Movilidad y Transporte: https://www.madrid.es/portales/munimadrid/es/Inicio/Movilidad-y-transportes/Madrid-Central-Zona-de-Bajas-Emisiones/Informacion-general/Madrid-Central-Informacion-General/?vgnextfmt=default&vgnextoid=a67cda4581f64610VgnVCM2000001f4a900aRCRD&vgnextchannel=0

Ayuntamiento de Madrid. (24 de Octubre de 2020). *EMT usará el Big Data de Moovit para mejorar la movilidad y el transporte público.* Obtenido de Diario de Madrid: https://diario.madrid.es/blog/notas-de-prensa/emt-usara-el-big-data-de-moovit-para-mejorar-la-movilidad-y-el-transporte-publico/

Azima, K. (2 de julio de 2018). *How do Transit Apps Know Bus and Train Arrival Times?* Obtenido de Medium: https://medium.com/@CommuterKate/how-do-transit-apps-know-bus-and-train-arrival-times-a3f15516487c

Belver, M., & R. Roces, P. (1 de octubre de 2019). *Almeida levanta la prohibición de entrar en Madrid Central a los vehículos con etiqueta C y al menos dos ocupantes.* Obtenido de El Mundo: https://www.elmundo.es/madrid/2019/09/30/5d91b34ffdddff34a78b4648.html

Berndt, D. J., & Clifford, J. (1994). *Using Dynamic Time Warping to Find Patterns in Time Series.* Nueva York: AAAI Technical Report. Obtenido de https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf

BiciMAD. (2020). *Qué es BiciMAD*. Obtenido de BiciMAD: https://www.bicimad.com/index.php?s=que

Borthakur, D., Sarma, J. S., Gray, J., Muthukkaruppan, K., Spiegelberg, N., Kuang, H., . . . Aiyer, A. (2011). *Apache Hadoop Goes Realtime at Facebook.* Atenas: Facebook Research. Obtenido de https://research.fb.com/publications/apache-hadoop-goes-realtime-at-facebook/

Breur, T. (2016). Statistical Power Analysis and the contemporary "crisis" in social sciences. *Journal of Marketing Analytics, 4*, 61-65. Obtenido de https://link.springer.com/content/pdf/10.1057/s41270-016-0001-3.pdf

Burkov, A. (2019). *The Hundred-Page Machine Learning Book.* Quebec: Andriy Burkov.

Cass, S. (6 de septiembre de 2019). *The Top Programming Languages 2019*. Obtenido de IEEE Spectrum.org: https://spectrum.ieee.org/computing/software/the-top-programming-languages-2019

Chamberlin, D. D. (2012). Early History of SQL. *IEEE Annals of the History of Computing*, 78-82. Obtenido de https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6359709

Codd, E. F. (1969). *A Relational Model of Data for Large Shared Data Banks.* San José: IBM Research Laboratory. Obtenido de https://cs.uwaterloo.ca/~david/cs848s14/codd-relational.pdf

Dedić, N., & Stanier, C. (2016). Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery. *ERP Future: International Conference on Enterprise Resource Planning Systems* (págs. 114-122). Hagenberg: Springer.

Dhawan, R., Hensley, R., Padhi, A., & Tschiesner, A. (2019). Mobility's second great inflection point. *McKinsey Quarterly*, 1-11. Obtenido de https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/mobilitys-second-great-inflection-point/es-es

Eclipse SUMO. (2020). *Simulation of Urban MObility*. Obtenido de SUMO Documentation: https://sumo.dlr.de/docs/

Evans, E. (12 de mayo de 2009). *NOSQL 2009*. Obtenido de Eric Evans's Weblog: https://web.archive.org/web/20110716174012/http://blog.sym-link.com/2009/05/12/nosql_2009.html

Fattah, J., Ezzine, L., Aman, Z., Moussami, H. E., & Lachhab, A. (2018). Forecasting of demand using ARIMA model. *International Journal of Engineering Business Management, 10*, 1-9. Obtenido de https://journals.sagepub.com/doi/pdf/10.1177/1847979018808673

Favarò, F. M., Nader, N., Eurich, S. O., Tripp, M., & Varadaraju, N. (20 de septiembre de 2017). Examining accident reports involving autonomous vehicles in California. (X. Hu, Ed.) *PLOS ONE*. Obtenido de https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0184952&type=printable

Fox, C. (2018). *Data Science for Transport.* Springer.

Ghahramani, Z. (18 de diciembre de 2019). *Uber AI in 2019: Advancing Mobility with Artificial Intelligence*. Obtenido de Uber Engineering: https://eng.uber.com/uber-ai-blog-2019/

GLH. (2018). *GLH to operate London's largest zero-emission cab fleet after eConnect Cars acquisition*. Obtenido de GLH : https://glh.co.uk/glh-to-operate-londons-largest-zero-emission-cab-fleet/

Hazarika, A. V., Ram, G. J., & Jain, E. (2017). Performance comparision of Hadoop and spark engine. *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)* (págs. 671–674). Palladam: IEEE. Obtenido de https://ieeexplore.ieee.org/document/8058263

HERE. (2020). *See Madrid in motion*. Obtenido de Urban Mobility Index - HERE Technologies: https://urbanmobilityindex.here.com/city/madrid/

Indra Sistemas. (20 de marzo de 2017). *Indra lidera el proyecto Transforming Transport, que utilizará el Big Data para mejorar la movilidad en Europa*. Obtenido de Indra Company : https://www.indracompany.com/sites/default/files/170320_np_indra_lidera_el_proyecto_transforming_transport_que_utilizara_el_big_data_para_mejorar_la_movilidad_en_europa.pdf

INE España. (2019). *Cifras oficiales de población resultantes de la revisión del Padrón municipal - Madrid*. Obtenido de Instituto Nacional de Estadística: https://www.ine.es/jaxiT3/Datos.htm?t=2881#!tabs-tabla

INE España. (2019). *Principales series desde 1971 - Población residente en España*. Obtenido de Instituto Nacional de Estadística: https://www.ine.es/jaxiT3/Datos.htm?t=31304#!tabs-tabla

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* Los Angeles: Springer. Obtenido de https://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

John Mashey. (25 de abril de 1990). *Big Data and the Next Wave of InfraStress.* Obtenido de The Advanced Computing System Association: https://static.usenix.org/event/usenix99/invited_talks/mashey.pdf

Jorrín, J., Zuil, M., Escudero, J., Martín, L., Learte, P., Rodríguez, L., . . . Hernández, A. (27 de Septiembre de 2019). *El Éxodo Urbano De España: La Nueva Despoblación del Siglo XXI*. Obtenido de El Confidencial: https://www.elconfidencial.com/economia/2019-09-27/exodo-urbano-espana-migraciones-provincias_2240119/

Kassambara, A. (2017). *Practical Guide To Cluster Analysis in R.* sthda.com.

Kitchin, R. (2014). The Real-Time City? Big Data and Smart Urbanism. *Geojournal*, 1-14. Obtenido de https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2289141

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity and Variety.* Stamford: META Group (Gartner en la actualidad). Obtenido de https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lith, A., & Mattsson, J. (2010). *Investigating storage solutions for large data: A comparison of well performing and scalable data storage.* Department of Computer Science and Engineering. Göteborg: Chalmers university of technology. Obtenido de http://publications.lib.chalmers.se/records/fulltext/123839.pdf

Liu, J., Sun, L., Chen, W., & Xiong, H. (agosto de 2016). Rebalancing Bike Sharing Systems: A Multi-source Data Smart Optimization. *roceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (págs. 1005–1014). Nueva York: Association of Computing Machinery.

MathWorks. (s.f.). *Forecast Multiplicative ARIMA Model*. Obtenido de MathWorks Documentation - Help Center: https://es.mathworks.com/help/econ/forecast-airline-passenger-counts.html

Mayor of London. (2019). *The Mayor's Ultra Low Emission Zone for London*. Obtenido de Mayor of London - London Government: https://www.london.gov.uk/what-we-do/environment/pollution-and-air-quality/mayors-ultra-low-emission-zone-london

Mayor of London. (21 de octubre de 2019). *ULEZ reduces 13,500 cars daily & cuts toxir air pollution by a third.* Obtenido de Mayor of London - UK Government: https://www.london.gov.uk/press-releases/mayoral/ulez-reduces-polluting-cars-by-13500-every-day

McQueen, B. (2017). Big data analytics for connected vehicles and smart cities. Norwood: Artech House.

Minsky, M., & Papert, S. (1969). A Review of "Perceptrons: An Introduction to Computational Geometry. *Information and Control, 17*, 501-522. Obtenido de https://pdf.sciencedirectassets.com/273276/1-s2.0-S0019995800X01222/1-s2.0-S0019995870904092/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEIv%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJIMEYCIQDw%2FXaNgZMyOO7omP2Z9w0Z8bFP%2BrReY2qMP2Y7lNzLuwIhAMKuprhB

Möller, T., Padhi, A., Pinner, D., & Tschiesner, A. (19 de diciembre de 2019). The future of mobility is at our doorstep. *McKinsey Quarterly*, 3-16. Obtenido de https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/the-future-of-mobility-is-at-our-doorstep

Moovit. (2020). *Leading MaaS Solutions that Meet Your Needs*. Obtenido de Moovit: https://moovit.com/maas-solutions/

Nuaimi, E. A., Neyadi, H. A., Mohamed, N., & Al-Jaroodi, J. (2015). Applications of big data to smart cities. *Journal of Internet Services and Applications*. Obtenido de https://link.springer.com/content/pdf/10.1186/s13174-015-0041-5.pdf

O'Mahony, E., & Shmoys, D. B. (2015). Data Analysis and Optimization for (Citi)Bike Sharing. *Twenty-Ninth AAAI Conference on Artificial Intelligence* (págs. 687-694). Austin: Cornell University.

Petrevska, B. (16 de mayo de 2017). Predicting tourism demand by A.R.I.M.A Models. *Economic Research-Ekonomska Istraživanja*, págs. 939-950. Obtenido de https://doi.org/10.1080/1331677X.2017.1314822

Portal OPENDATA - EMT. (31 de marzo de 2017). *Modelo de datos de la información de uso del servicio BiciMAD.* Obtenido de OPENDATA - EMT- BiciMAD: https://opendata.emtmadrid.es/Documentos/Servicios-y-estructuras-Bicimad-V1-1.aspx

PTV Group. (2020). *PTV Vissim*. Obtenido de PTV Products - Products: https://www.ptvgroup.com/en/solutions/products/ptv-vissim/

Python.org. (2020). *Success Stories - Business*. Obtenido de Python Software Foundation: https://www.python.org/success-stories/category/business/

RACE. (26 de febrero de 2019). *Zona azul y verde: ¿cuánto cuesta aparcar en tu ciudad?* Obtenido de Real Automóvil Club de España: https://www.race.es/zona-azul-verde-precio-ciudades

Ratanamahatana, C. A., & Keogh, E. (2004). Making Time-series Classification More Accurate Using Learned Constraints. *SIAM International Conference on Data Mining* (págs. 11-22). Lake Buena Vista: SIAM Journal on Scientific Computing (ISIC). Obtenido de https://www.cs.ucr.edu/~eamonn/RATANAMC.pdf

Renfe. (4 de junio de 2020). *Renfe pone en marcha un proyecto piloto para controlar el aforo en las estaciones de Cercanías en tiempo real*. Obtenido de Renfe - Sala de Prensa: https://saladeprensa.renfe.com/renfe-pone-en-marcha-un-proyecto-piloto-para-controlar-el-aforo-en-las-estaciones-de-cercanias-en-tiempo-real/

RStudio. (2020). *RStudio - Inicio*. Obtenido de RStudio: https://rstudio.com/

Sánchez, J. L., & Díaz, B. (19 de febrero de 2019). *Accenture y Metro de Madrid equilibran eficiencia energética y confort gracias a la Inteligencia Artificial*. Obtenido de Accenture - Sala de Prensa: https://www.accenture.com/es-es/company-news-release-accenture-metro-madrid

Singhvi, D., Singhvi, S., Frazier, P. I., Henderson, S. G., Mahony, E. O., Shmoys, D. B., & B.Woodard, D. (2015). Predicting Bike Usage for New York City's Bike Sharing System. *2015 AAAI Workshop* (págs. 110-114). Palo Alto: Cornell University. Obtenido de https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/viewFile/10115/10185

Stahl, F., May, D., Mills, H., Bramer, M., & Gaber, M. M. (2015). A Scalable Expressive Ensemble Learning Using Random Prism: A MapReduce Approach. En *Transactions on Large-Scale Data- and Knowledge-Centered Systems XX* (págs. pp 90-107). Berlin: Springer Berlin Heidelberg. Obtenido de https://link.springer.com/chapter/10.1007%2F978-3-662-46703-9_4

Strauch, C. (2009). *NoSQL Databases.* Stuttgart: Hochschule der medien Stuttgart. Obtenido de https://www.christof-strauch.de/nosqldbs.pdf

Tableau Software. (2020). *Governed self-service analytics at scale with Tableau Server*. Obtenido de Tableau - Products: https://www.tableau.com/products/server

Tableau Software. (2020). *Travel and Transportation Analytics*. Obtenido de Tableau - Our Solutions: https://www.tableau.com/solutions/travel-and-transportation-analytics#reveal

Templeton, B. (6 de mayo de 2019). *Elon Musk's War On LIDAR: Who Is Right And Why Do They Think That?* Obtenido de Forbes: https://www.forbes.com/sites/bradtempleton/2019/05/06/elon-musks-war-on-lidar-who-is-right-and-why-do-they-think-that/#1e9d10d2a3bd

Tian, Y., Pei, K., Jana, S., & Ray, B. (2018). DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars. *2018 ACM/IEEE 40th International Conference on Software Engineering* (págs. 303-314). Gotemburgo: Association for Computing Machinery. Obtenido de https://dl.acm.org/doi/pdf/10.1145/3180155.3180220

Transforming Transport. (2018). *AUSOL Load Balancing Pilot*. Obtenido de transformingtransport.eu: https://transformingtransport.eu/transport-domains/ausol-load-balancing-pilot

Transport for London. (2003). *Driving - Congestion Charge.* Obtenido de TFL - UK government: https://tfl.gov.uk/modes/driving/congestion-charge

Transport for London. (30 de julio de 2018). *Press Release - Cycle hire scheme celebrates best ever month of hires*. Obtenido de TFL - UK Government: https://tfl.gov.uk/info-for/media/press-releases/2018/july/cycle-hire-scheme-celebrates-best-ever-month-of-hires

Transport for London. (2019). *TFL- UK Government.* Obtenido de ULEZ - Discounts and exemptions: https://tfl.gov.uk/modes/driving/ultra-low-emission-zone/discounts-and-exemptions

Verendel, V., & Yeh, S. (2019). Measuring Trafc in Cities Through a Large-Scale Online Platform. *Journal of Big Data Analytics in Transportation, 1*, 161–173. Obtenido de https://link.springer.com/content/pdf/10.1007/s42421-019-00007-7.pdf

Wang, S. J., & Moriarty, P. (2018). Big Data for Sustainable Urban Transport. En S. J. Wang, & P. Moriarty, *Big data for urban sustainability* (págs. 81-103). Springer.

Waymo. (10 de octubre de 2018). *Waymo Celebrates 10 Million Miles of Self-Driving*. Obtenido de Waymo's Technology: https://waymo.com/tech/

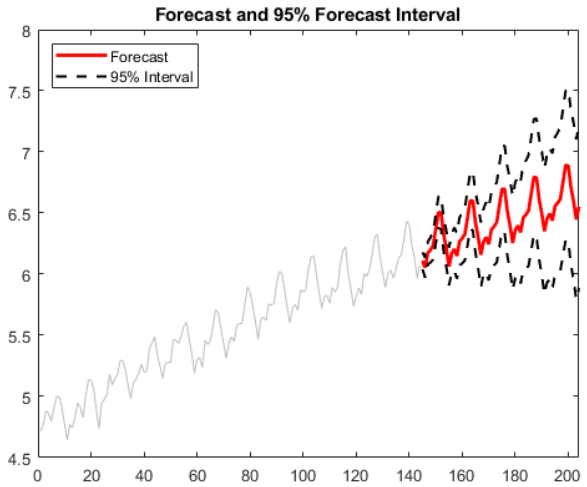
Whitehead, F. (14 de octubre de 2010). *London bike hire scheme on road to be only public transport system in profit*. Obtenido de The Guardian: https://www.theguardian.com/environment/green-living-blog/2010/oct/13/london-bike-hire-profit

ANEXOS

ANEXO I. Futuras líneas de investigación: Modelos SARIMA, regresión lineal climática y reglas de asociación

Estimación de demanda mediante un modelo estadístico ARIMA

Esta regresión es un método estadístico para series temporales y generalmente, no es considerado un método de aprendizaje automático. Es una generalización del modelo ARMA. El modelo ARIMA parte de la idea de que se pueden utilizar valores pasados de una serie temporal para predecir valores futuros de la misma, por lo que es un modelo autorregresivo donde la variable dependiente y explicativa son la misma, pero están en distinto orden temporal. La variable dependiente estará en el momento t, y la variable explicativa en el momento t-1. Resulta de extraordinaria utilidad en la predicción de demanda futura y se ha utilizado en numerosos estudios de demanda de electricidad o de turismo (Petrevska, 2017). El modelo ARIMA siempre está indicado por la forma ARIMA (P, D, Q) donde P es el número de intervalos de tiempo del modelo, D es el grado de diferenciación o el número de veces que se han restado valores pasados a los datos, y Q es el orden del modelo de promedio móvil (Fattah, Ezzine, Aman, Moussami, & Lachhab, 2018).

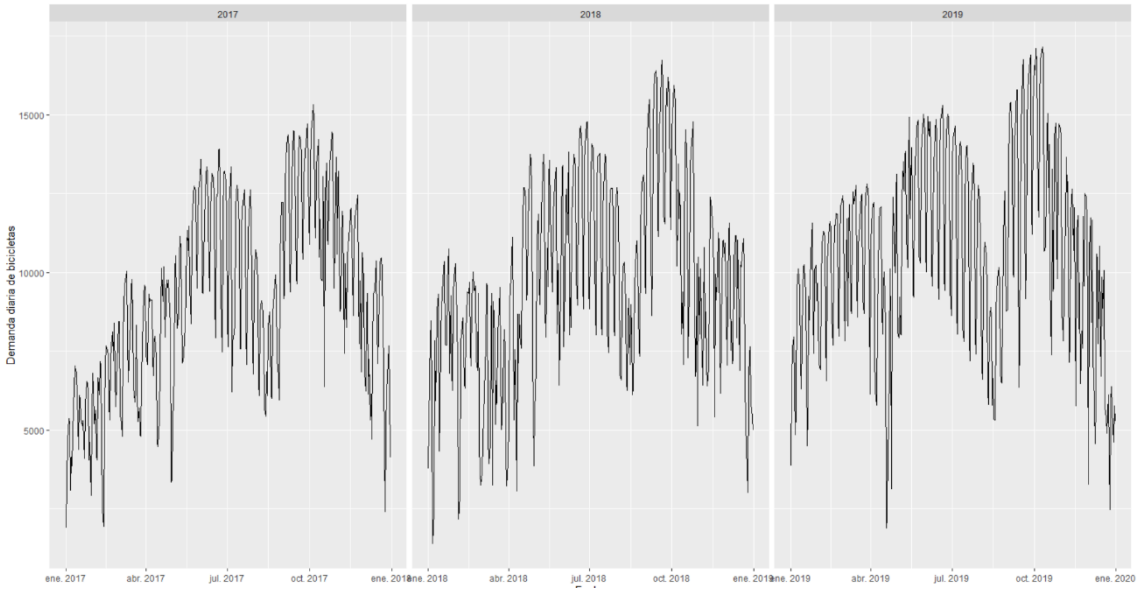


Modelo ARIMA, predicción e intervalo de confianza (MathWorks, s.f.)

Se ha realizado un pequeño estudio ARIMA. Se ha obtenido la serie temporal consolidada del uso o demanda diaria mediante la siguiente URL:

https://datos.madrid.es/egob/catalogo/213155-1-bicimad-usos-usuarios.xls

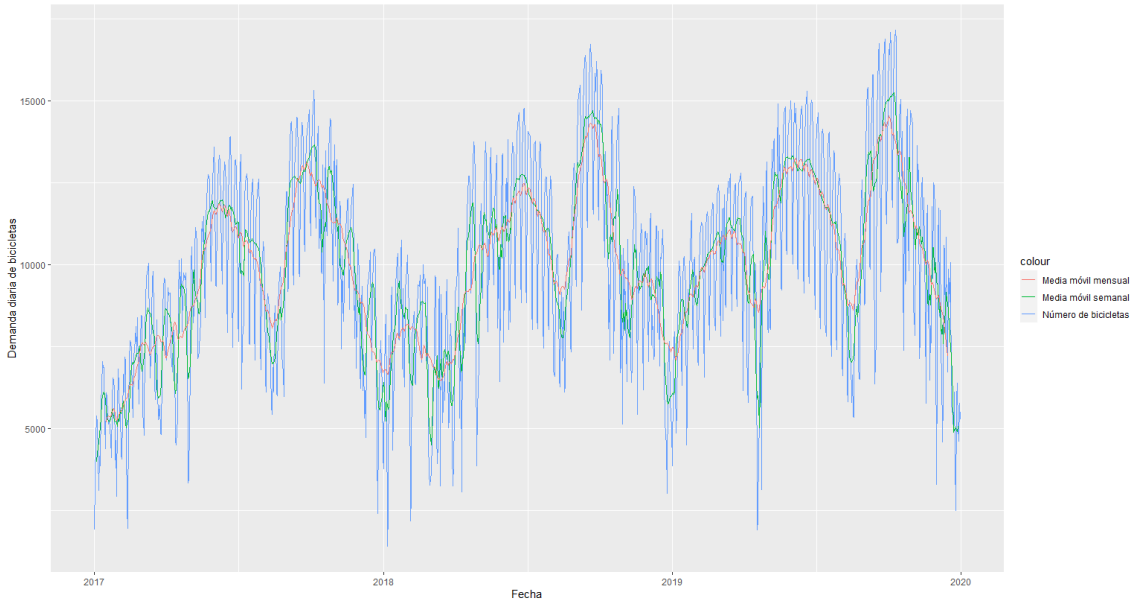
Esta serie temporal se puede conseguir de la primera base de datos utilizada, aunque únicamente para el año 2018. Se utilizan las series temporales del año 2017, 2018 y 2019, cuyo gráfico se puede observar a continuación.



BiciMAD – Serie temporal de la demanda 2017, 2018 y 2019

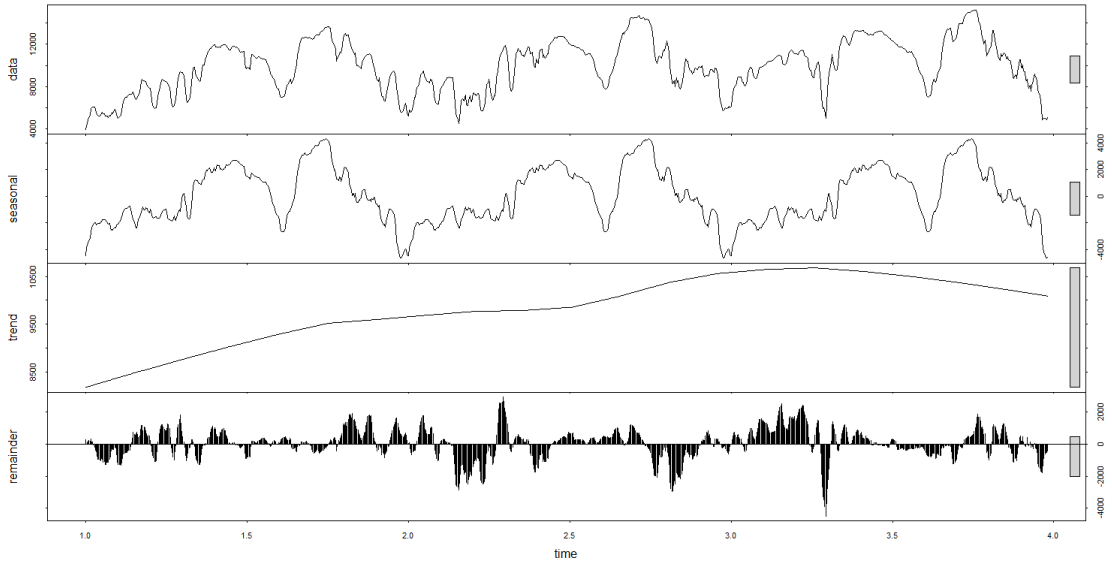
Como se observa, esta serie temporal tiene un componente estacionario (debido a que el comportamiento se repite año tras año) y un componente estacional (ya que se nota un ligero aumento año tras año). Para realizar un modelo ARIMA, debemos calcular las medias móviles y comprobar si la serie temporal es estacionaria mediante la prueba de Dickey-Fuller. Se calculan las medias móviles y se realiza la prueba, la cual resulta en un p-valor de 0,0397, inferior a 0,05 por lo que se rechaza la hipótesis nula y se afirma que la serie es estacionaria.

Esto también nos indica que el sistema se comporta de manera similar a lo largo del tiempo en demanda total, especialmente al año anterior, lo cual, indirectamente, da fuerza al argumento discutido en la conclusión de este TFG: analizar agrupaciones de estaciones por su comportamiento cada mes del año, con el fin de balancear mejor el sistema cada mes al año siguiente. No obstante, habría que comprobar si las estaciones de septiembre de 2019 se comportan de manera similar a septiembre de 2018, ya que esto verificaría nuestra teoría.



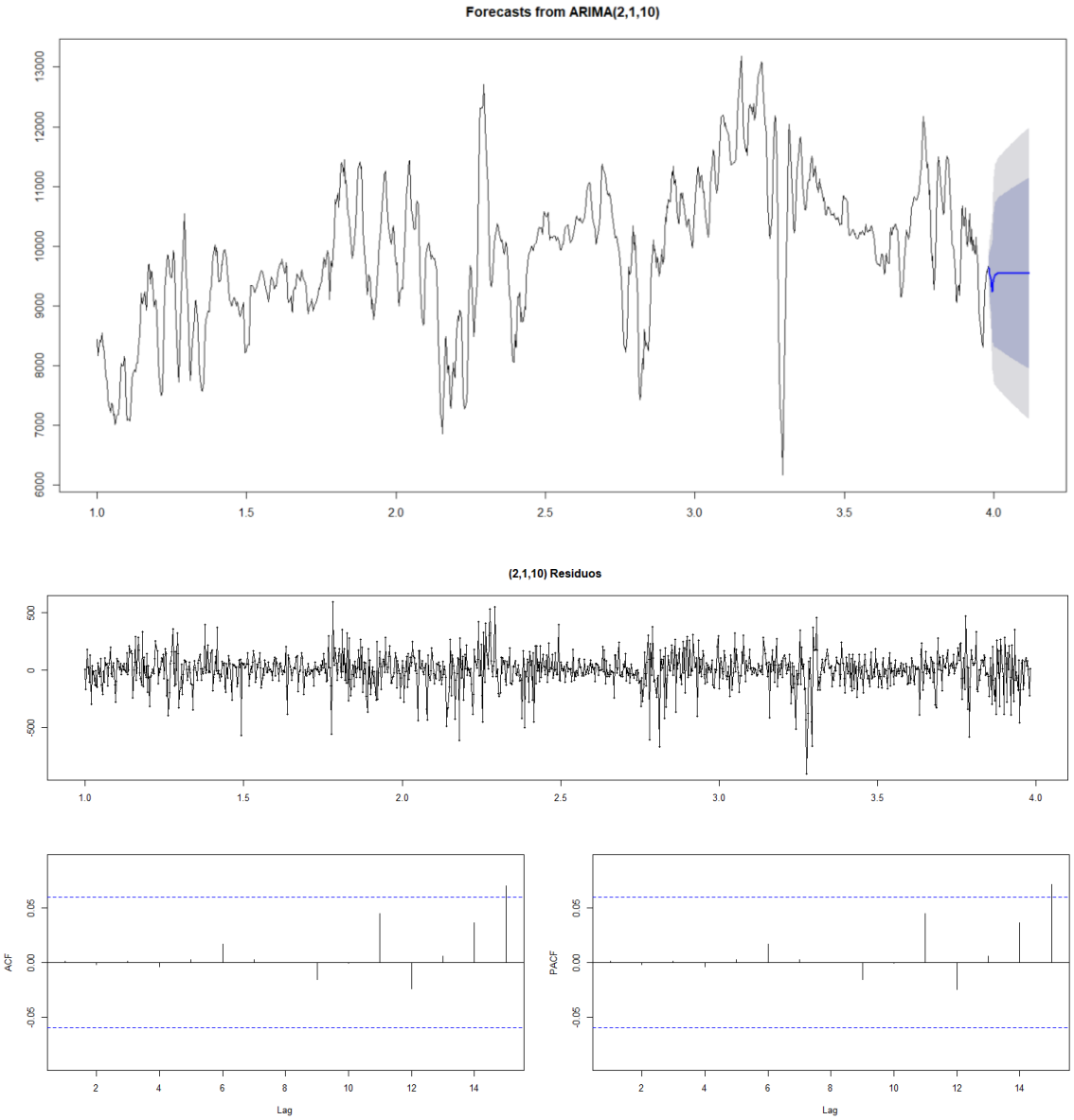
BiciMAD – Serie temporal de la demanda 2017, 2018 y 2019 y medias móviles

Una vez realizado esto, se descompone la serie temporal en estacionalidad, tendencia y lo que sobra tras haberle quitado estos dos componentes.



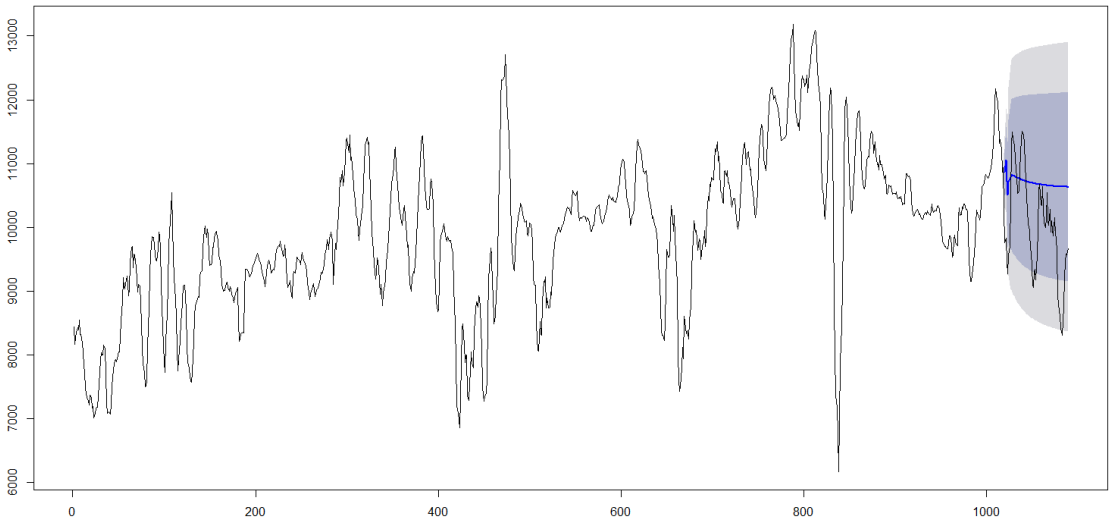
BiciMAD – Descomposición de la serie temporal

Se estima el mejor modelo ARIMA en ARIMA (2,1,3). Sin embargo, sus autocorrelogramas no son aceptables para realizar predicciones debido al exceso de retardos, por lo que se realiza un ARIMA (2,1,10) como modelo final, para el cual se realiza una predicción a 50 días.



BiciMAD – Modelo ARIMA (2, 1, 10), distribución de residuos y sus autocorrelogramas

De la misma manera, se realiza la predicción con ARIMA (2, 1, 10) sobre la misma serie temporales, pero para una fecha anterior, de tal modo que se puede comparar la eficacia del modelo con lo que realmente pasó. Para este caso, el modelo logra captar la mayoría de los datos históricos dentro del intervalo de confianza del 95% y otros pocos dentro del intervalo de confianza del 80%.



BiciMAD – Modelo ARIMA (2, 1, 10) para una fecha anterior de la serie, predicción y representación real de estos

Se ha descartado realizar un modelo SARIMA debido a la naturaleza de los datos y a limitaciones de carácter computacional. Un modelo SARIMA, al contrario que uno ARIMA, tiene en cuenta la estacionalidad de la serie temporal y se ajusta más a la realidad. No obstante, queda demostrado que los modelos ARIMA son útiles para medir las tendencias que tendrá un servicio como BiciMAD en el futuro, con el fin de estar mejor preparados ante una gran demanda.

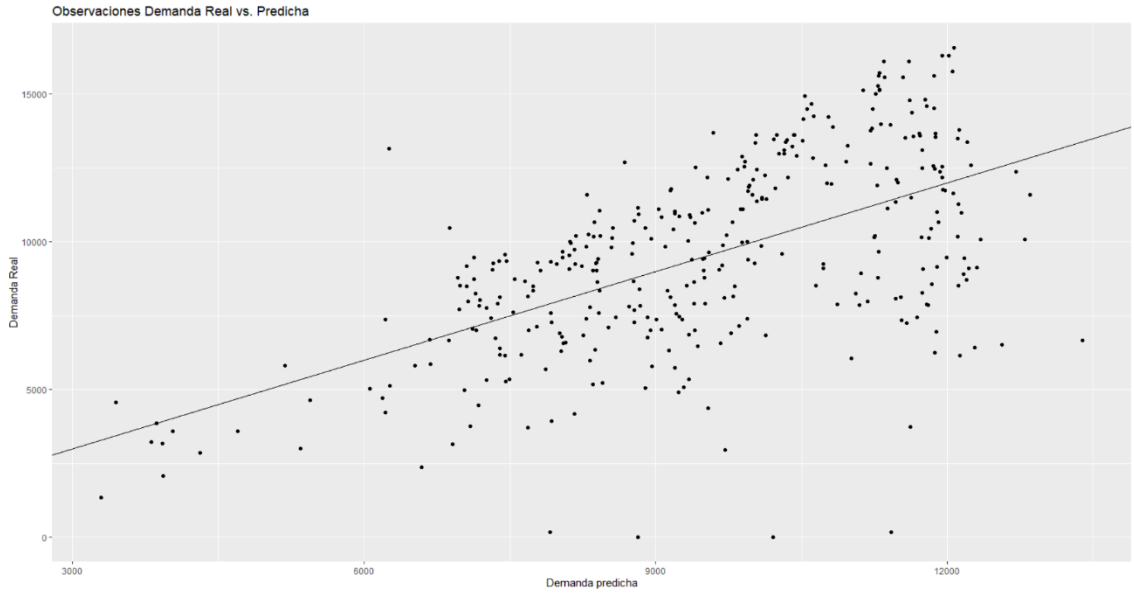
Estimación de demanda mediante un modelo de regresión lineal

Se ha construido el siguiente modelo de forma muy genérica, discriminando a la mayoría de las variables iniciales. Las variables que más influyen al modelo son el mes, la temperatura, la cantidad de lluvia y la racha de viento. De la misma manera, este estudio propuesto sobre la demanda está basado en un estudio previo sobre CitiBike que realiza regresiones lineales sobre la demanda de CitiBike y explora como esta es afectada por variables climáticas, población y uso de taxis (Singhvi, y otros, 2015).

```
lm(formula = Demanda ~ Month + Tmean + Rain_amount + Wind_Gust,
    data = datos_modelo)

Residuals:
     Min       1Q   Median       3Q      Max
-11019.7  -1685.4    339.8   1865.2   4481.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8030.63     632.16  12.703  < 2e-16 ***
Month          85.97      43.43   1.980  0.04859 *
Tmean         174.37      18.61   9.370  < 2e-16 ***
Rain_amount  -181.66      39.59  -4.588 6.37e-06 ***
Wind_Gust    -162.20      50.95  -3.183  0.00159 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2548 on 329 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared:  0.3825,    Adjusted R-squared:  0.375
F-statistic: 50.95 on 4 and 329 DF,  p-value: < 2.2e-16
```

Modelo de Regresión lineal realizado con la demanda y las variables climáticas

Se debe de tener en cuenta que este modelo se ha construido con dos bases de datos independientes y sin relación alguna, más allá de ser el sistema de BiciMAD un medio de transporte muy expuesto al clima, y que no se ha añadido otra variable de la misma base de datos de viajes más que el recuento o demanda.



Regresión lineal: Gráfica de valores reales contra los predichos por el modelo

Aun así, esta regresión lineal tiene varios fallos debido a la naturaleza de los datos, como hemos comentado con la medición de la variable de precipitación en el apartado 4.4.3 a modo de ejemplo. En el punto 4.5, análisis climático, la mayoría de estar variables no tienen una relación lineal, por lo que deberían realizarse transformaciones no lineales sobre estas variables con el uso de logaritmos o exponenciales. Otra opción es realizar un modelo no lineal desde el inicio. A modo de mejora, el R^2 ajustado de 0,375 sería mayor si tuviésemos datos climáticos por hora, algo de lo que no se dispone. En el anterior

gráfico, un modelo perfecto sería aquel donde la demanda agregada diaria sea igual o muy pareja a aquella demanda predicha, donde x = y. Posteriormente, el modelo debería validarse mediante la prueba de White de la heterocedasticidad, una prueba de multicolinealidad y otra de distribución de los residuos.

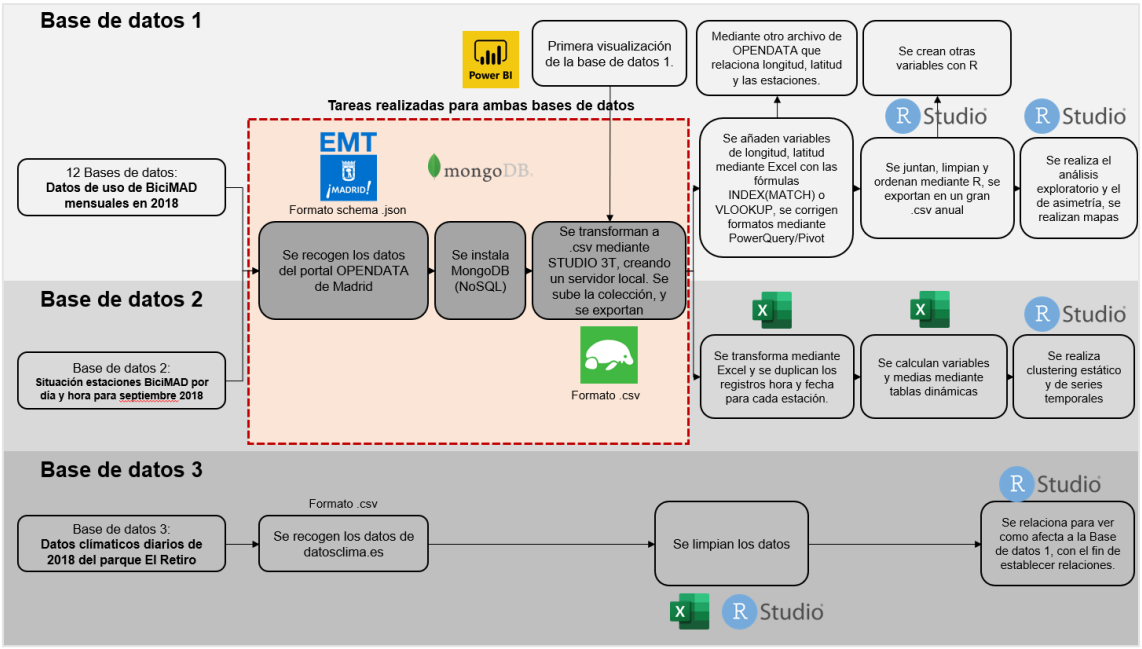
Demanda: Gráfica de valores reales rojos frente a los predichos por el modelo en azul

En este gráfico el azul determina la demanda predicha mientras que el rojo determina la demanda agregada diaria real. Nuestro modelo podría llegar a predecir de una manera muy ajustada si tuviésemos mayores datos climáticos por hora. Se ha construido un modelo muy general por lo que este análisis se establece a modo de sugerencia, para determinar la cantidad de empleados necesarios por día.

Aprendizaje no supervisado: Reglas de asociación

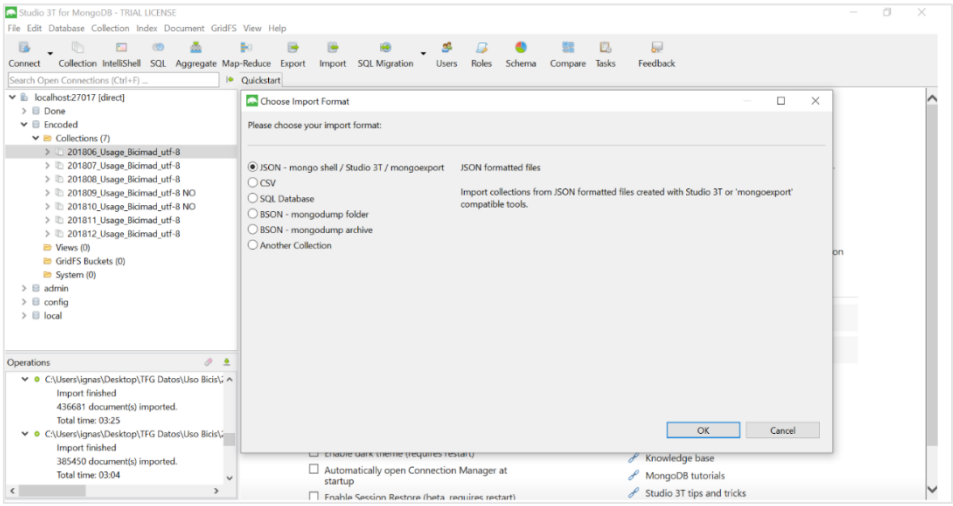
Se ha pensado en aplicar reglas de asociación con la primera base de datos. Este algoritmo de carácter no supervisado realiza una búsqueda de relaciones y podría ser una gran idea para saber, con que probabilidad se dan las distintas rutas saliendo de una determinada estación en un sistema de transporte. Esto no solamente permite conocer los flujos principales de la ciudad (lo cual es descrito por el apartado 4.5.5, las rutas más populares en el caso BiciMAD), sino que nos da un mapa de probabilidades de todas las estaciones para un determinado día del año. La idea es centrarse en las relaciones más fuertes para cada estación, de tal forma que los empleados puedan comprender cómo se dan los flujos de manera rápida y anticiparse a los recorridos que se deben efectuar.

ANEXO II. Transformación de los datos



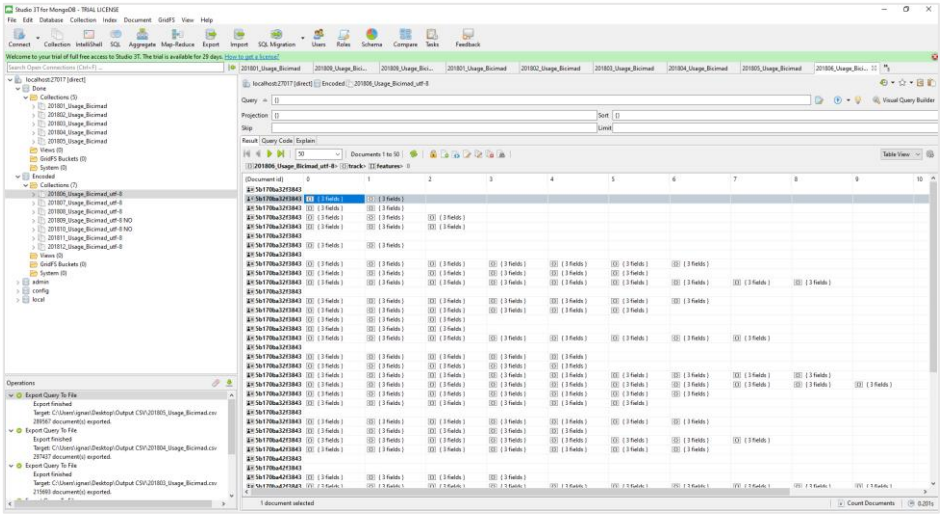
Flujos de transformaciones realizadas a cada base de datos

Se ha preparado el siguiente diagrama con el fin de ver la transformación de los datos realizada. Como se puede observar por la imagen, la base de datos 1 y 2 han sido trabajadas extensivamente, mientras que la base de datos 3 ha sido aplicada casi sin transformación alguna. A continuación, se dejan algunas imágenes de los procesos seguidos.



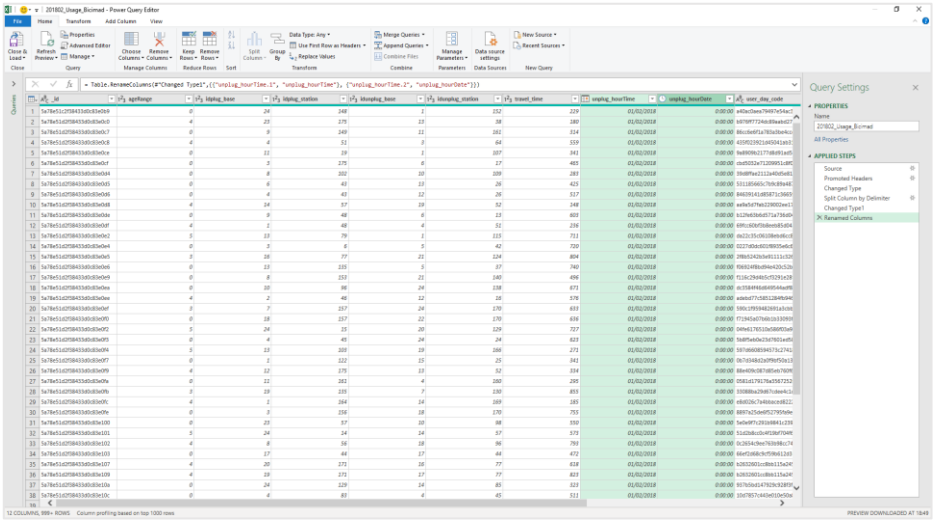
Importación de bases de datos mensuales mediante STUDIO 3T

Esta imagen muestra los datos de la primera base de datos siendo importados al programa STUDIO3T. Se importaron como archivos esquema MongoDB .json para ser exportados como .csv posteriormente.

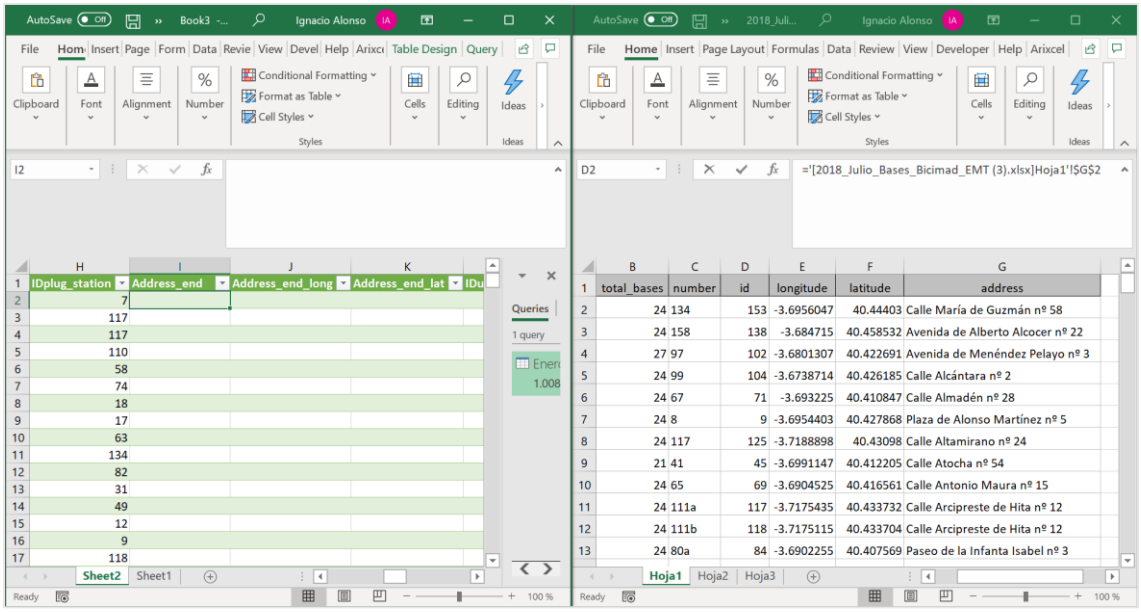


Algunos registros de la variable "track" de geolocalización

Esta imagen muestra la variable "track", que mide la geolocalización de la bicicleta cada 30 segundos, además de otras variables como la velocidad de esta. El problema es que, para viajes de mayor duración, la variable "track" realiza mediciones sin parar. Por lo que, como se puede ver en la imagen, cada viaje o registro tiene un número de variables "track" muy diferentes. En la imagen se pueden ver viajes de 2 mediciones a 9 mediciones.

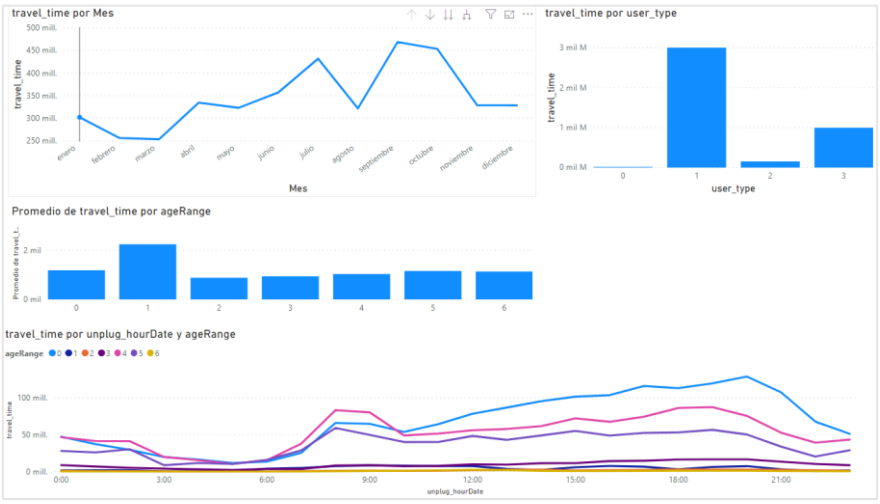


Transformación y separación del formato fecha y hora en Power Query Editor



Proceso VLOOKUP/INDEX(MATCH) para añadir las variables de latitud y longitud y nombre de la estación a cada base de datos mensual.

También se muestran capturas de las diferentes transformaciones en Excel. Se realizaron transformaciones para separar la fecha y hora (que estaban juntas en los archivos originales .json y .csv) mediante PowerQuery y se añadieron las variables de latitud y longitud para las 12 bases de datos mensuales, mediante el uso de un archivo que relacionaba la i.d de la estación y su ubicación. En la transformación de la primera base de datos, el mapeado se realiza mediante las funciones INDEX(MATCH) o VLOOKUP (BUSCARV en castellano).

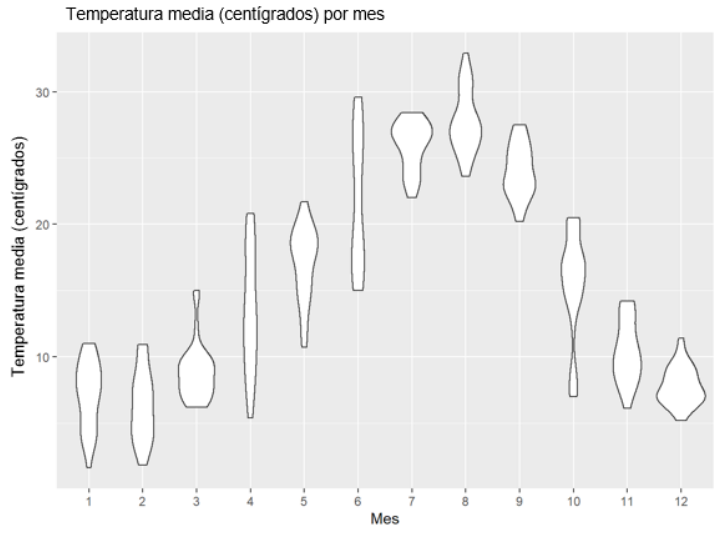


Importación de bases de datos mensuales mediante STUDIO 3T

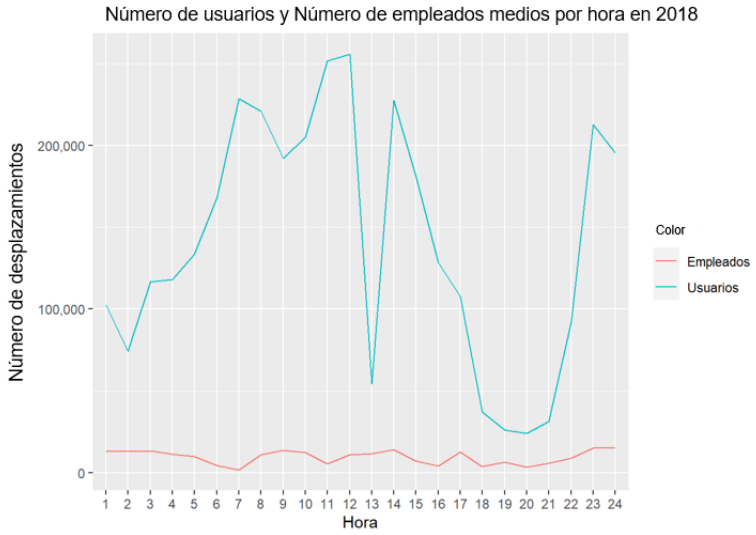
Por último, se muestra el panel inicial realizado en PowerBI, como medida de control para ver si los datos daban relaciones que tuviesen sentido. Se realizó con la variable duración del recorrido o travel_time.

ANEXO III. Otras gráficas para el análisis descriptivo

En este anexo se incluyen otras gráficas realizadas e interesantes que no se han incluido finalmente en el caso práctico. En este gráfico se puede ver cómo cambia la temperatura por mes, la temperatura y la demanda de bicicletas están muy correlacionados.

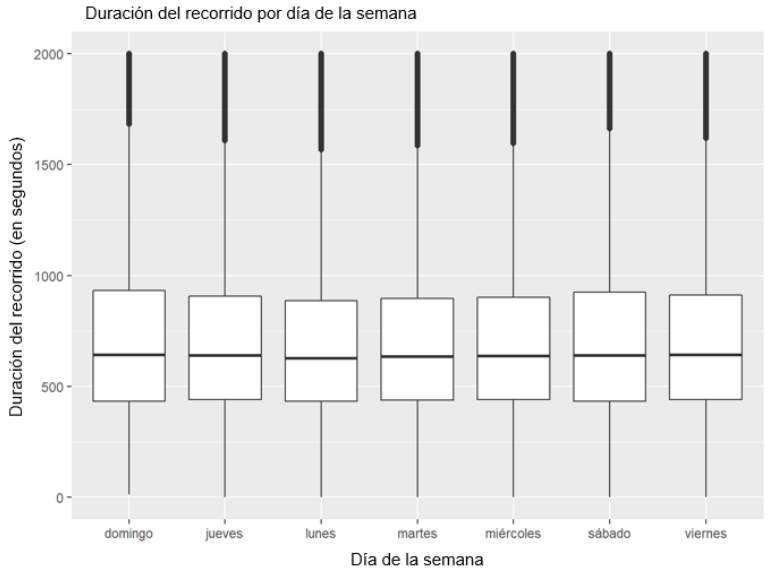


Temperatura media, rango y densidad (centígrados) por mes del año



Recuento del número de desplazamientos por hora y tipo de usuario

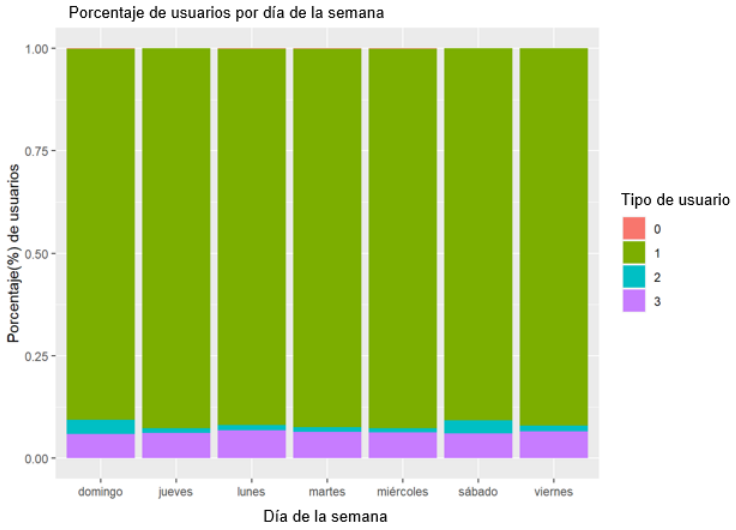
El número de usuarios es muy superior al de empleados. El número de usuarios crece rápidamente por la mañana y tarde mientras que el de los empleados se mantiene constante. Destaca el menor número de desplazamientos por parte de los usuarios antes del almuerzo sobre las 13:00 y en el rango de 18:00 a 21:00.



Duración del recorrido por día de la semana

En este gráfico destaca la mayor duración de los trayectos el fin de semana, posiblemente debido al mayor número de usuarios ocasionales.

A continuación, se comprueba que, efectivamente, hay más usuarios del tipo dos en los fines de semana. También se corrobora el dominio del tipo de usuario uno o abonado. Los empleados se comportan de manera estable, estando algo más presentes en los días laborables.



Porcentaje de usuarios por día de la semana

ANEXO IV. Detalle: tablas de asimetría por estación del año en RStudio

Asimetría de las diez estaciones con mayor y menor asimetría en invierno

```
## # A tibble: 10 x 4
##    Estación_Origen             Salidas   Entradas   Asimetría_total
##    <fct>                         <int>    <int>           <int>
##  1 Paseo de la Chopera nº14       7503    10082            2579
##  2 Paseo de la Florida nº8        5119     7349            2230
##  3 C/Cerro de la Plata nº2        3183     4673            1490
##  4 Paseo Santa María Cabeza nº58  7815     9054            1239
##  5 Plaza de la Cebada nº16        8971     9934             963
##  6 Paseo de los Olmos nº28        5172     6071             899
##  7 C/Jaime el Conquistador nº30   8485     9341             856
##  8 Paseo de las Delicias nº92-94  4289     4925             636
##  9 C/Fernando el Católico nº19    7273     7902             629
## 10 C/Palos de la Frontera nº40    5272     5888             616
```

```
## # A tibble: 10 x 4
##    Estación_Origen            Salidas   Entradas    Asimetría_total
##    <fct>                        <int>    <int>           <int>
##  1 Paseo Castellana nº67         3663     2802            -861
##  2 Paseo de la Castellana nº4    2314     1561            -753
##  3 C/Carlos III nº1              4010     3403            -607
##  4 C/María de Guzmán nº58        3227     2656            -571
##  5 C/Valencia nº1                9439     8905            -534
##  6 C/Sor Ángela de la Cruz nº2   2475     1966            -509
##  7 Paseo Yeserías nº45           2270     1764            -506
##  8 Avda. de Menéndez Pelayo nº3  3111     2616            -495
##  9 Puerta del Sol nº1 - B        3386     2893            -493
## 10 C/Serrano nº34 - A            1888     1425            -463
```

Asimetría de las diez estaciones con mayor y menor asimetría en primavera

```
## # A tibble: 10 x 4
##    Estación_Origen            Salidas   Entradas    Asimetría_total
##    <fct>                         <int>    <int>            <int>
##  1 Paseo de la Chopera nº14      11036    14738             3702
##  2 Paseo de la Florida nº8        7206    10074             2868
##  3 C/Cerro de la Plata nº2        4679     6528             1849
##  4 Paseo Santa María Cabeza nº58  9947    11460             1513
##  5 Paseo de los Olmos nº28        6746     8074             1328
##  6 Plaza de la Cebada nº16       12200    13403             1203
##  7 Paseo de las Delicias nº92-94  6215     7111              896
##  8 C/Jaime el Conquistador nº30  11022    11816              794
##  9 Avda. de Menéndez Pelayo nº38  8501     9201              700
## 10 C/Fernando el Católico nº19    9328     9978              650
```

```
## # A tibble: 10 x 4
##    Estación_Origen            Salidas   Entradas    Asimetría_total
##    <fct>                         <int>    <int>            <int>
##  1 Paseo Castellana nº67          4482     3465            -1017
##  2 Paseo Castellana nº4           3209     2265             -944
##  3 C/Serrano nº34 - A             2702     1975             -727
##  4 Avda. de Menéndez Pelayo nº3   4293     3571             -722
##  5 C/Carlos III nº1               4824     4133             -691
##  6 Plaza San Juan de la Cruz nº11 3380     2699             -681
##  7 Paseo Castellana nº42          4993     4335             -658
##  8 C/Goya nº18                    6113     5503             -610
##  9 C/María de Guzmán nº58         4222     3627             -595
## 10 Puerta del Sol nº1 - B         4866     4288             -578
```

Asimetría de las diez estaciones con mayor y menor asimetría en verano

```
## # A tibble: 10 x 4
##    Estación_Origen            Salidas   Entradas    Asimetría_total
##    <fct>                        <int>    <int>            <int>
##  1 Paseo de la Chopera nº14      12746    17356            4610
##  2 Paseo de la Florida nº8        8914    11620            2706
##  3 Calle Cerro de la Plata nº2    5625     8125            2500
##  4 Paseo Santa María Cabeza nº58 11057    12910            1853
##  5 Paseo de los Olmos nº28        7939     9331            1392
##  6 C/Jaime el Conquistador nº30  13780    15047            1267
##  7 Plaza de la Cebada nº16       12955    14023            1068
##  8 Paseo de las Delicias nº92-94  7544     8597            1053
##  9 C/Fernando el Católico nº19   10366    11060             694
## 10 Plaza de San Francisco nº5     8111     8801             690
```

```
## # A tibble: 10 x 4
##    Estación_Origen              Salidas  Entradas   Asimetría_total
##    <fct>                          <int>    <int>            <int>
##  1 Paseo Castellana nº67           5525     4225            -1300
##  2 Paseo Castellana nº4            3715     2664            -1051
##  3 Avda. de Menéndez Pelayo nº3    4889     3949             -940
##  4 C/Carlos III nº1                5689     4789             -900
##  5 C/Serrano nº34 - A              3150     2358             -792
##  6 Puerta del Sol nº1 - B          5654     4867             -787
##  7 C/Serrano nº34 - B              3579     2813             -766
##  8 Paseo de la Castellana nº42     5614     4890             -724
##  9 C/Sor Ángela de la Cruz nº2     4127     3424             -703
## 10 C/Serrano nº54                  5965     5282             -683
```

Asimetría de las diez estaciones con mayor y menor asimetría en otoño

```
## # A tibble: 10 x 4
##    Estación_Origen             Salidas  Entradas   Asimetría_total
##    <fct>                          <int>    <int>           <int>
##  1 Paseo de la Chopera nº14       10830    15328            4498
##  2 Paseo de la Florida nº8         7377    10100            2723
##  3 Paseo Santa María Cabeza nº58  10813    13019            2206
##  4 C/Cerro de la Plata nº2         5363     7349            1986
##  5 C/Jaime el Conquistador nº30   12842    14136            1294
##  6 Plaza de la Cebada nº16        12287    13489            1202
##  7 Paseo de los Olmos nº28         7110     8191            1081
##  8 C/Fernando el Católico nº19    10508    11384             876
##  9 Paseo de las Delicias nº92-94   7100     7880             780
## 10 Avda. de Menéndez Pelayo nº38   9056     9637             581
```

```
## # A tibble: 10 x 4
##    Estación_Origen             Salidas  Entradas   Asimetría_total
##    <fct>                          <int>    <int>           <int>
##  1 Paseo Castellana nº67           4948     3876           -1072
##  2 C/Serrano nº34 - A              3082     2080           -1002
##  3 Avda. de Menéndez Pelayo nº3    4340     3345            -995
##  4 Paseo de la Castellana nº4      3228     2278            -950
##  5 C/Carlos III nº1                5559     4788            -771
##  6 C/Serrano nº34 - B              3358     2627            -731
##  7 C/Sor Ángela de la Cruz nº2     3958     3234            -724
##  8 C/Pavía nº6                     4083     3382            -701
##  9 C/María de Guzmán nº58          5154     4463            -691
## 10 Plaza de España - A             4942     4348            -594
```

ANEXO V. Código R, descarga de datos, análisis e informe en Google Drive

Todo el análisis realizado sobre el sistema de movilidad urbana BiciMAD se puede descargar desde esta carpeta Google Drive, los datos e informes en la carpeta se pueden utilizar libremente para posteriores estudios.

URL: Google Drive