# 1 Descripción del método de predicción

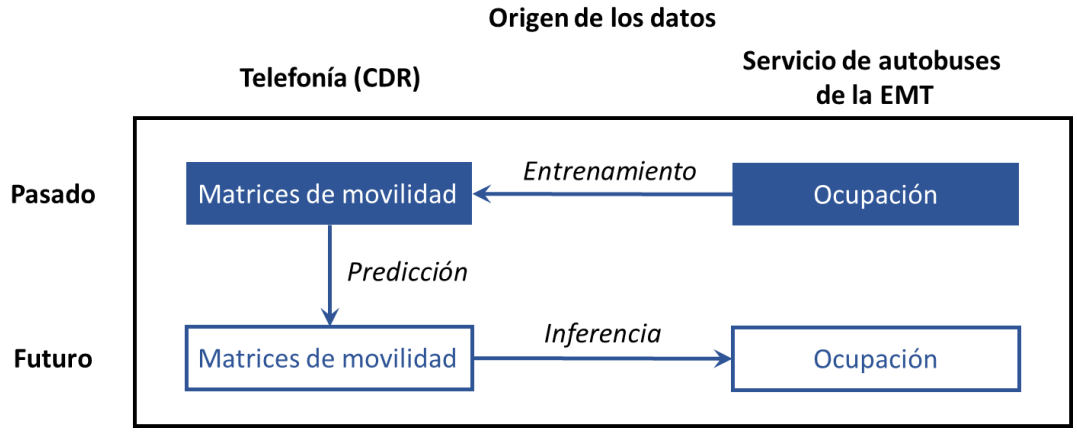
## 1.1 Estructura del método de predicción

El modelo que proponemos se apoya no sólo en los datos del servicio de autobuses de la EMT sino también en las matrices de movilidad generadas por Kido Dynamics a partir de los datos de telefonía móvil. Esto nos permite disponer de una imagen completa de los patrones de movilidad de toda la región, a cada hora del día y para todos los días del año, para entender qué fracción de esa movilidad tiene lugar a través del sistema de autobuses públicos.

Dada la exhaustividad y completitud de los datos de movilidad tanto en tiempo como espacio, es de esperar que las predicciones o proyecciones a futuro sean numéricamente más estables con reducidos márgenes de error numérico, centrando toda fuente de incertidumbre a la propia naturaleza inherente del comportamiento humano.

Bajo esta premisa, nuestro método se basa en tres pasos:

1. Entrenamiento con los datos facilitados por la EMT para cruzar ocupación de autobús por tramo con la matriz origen-destino para el histórico de datos a toro pasado.
2. Predicción a futuro de la matriz origen-destino.
3. Inferencia de las tablas de ocupación a futuro a partir de la predicción de la matriz de movilidad y la relación entre ocupación y movilidad entrenada en el primer paso.




La predicción puede parametrizarse al introducir en el proceso de entrenamiento e inferencia diferentes categorías de día:

- Día de la semana: LMMXVSD
- Tipo de día: Laborable / Festivo / Previo
- Meteorología: Soleado / Nublado / Lluvioso
- Eventos frecuentes: Bernabéu / IFEMA / etc.
- Movilidad excepcional: Matrices de movilidad modificadas

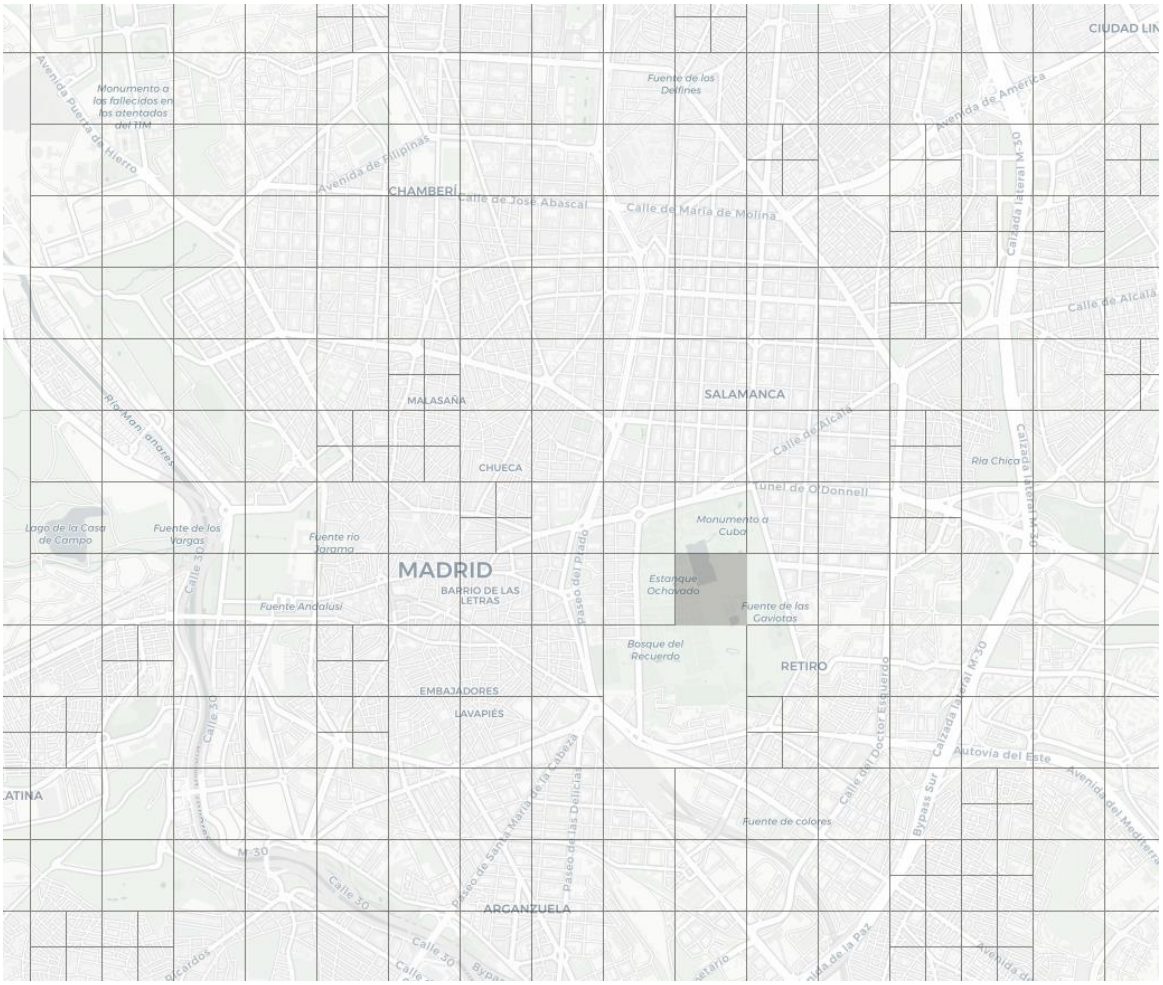
Es de esperar que estos parámetros afecten tanto a los patrones de movilidad generales como a las fracciones por cada modo, y que por tanto deban ser tenidos en cuenta tanto a nivel de entrenamiento como de predicción.

## 1.2 Fuentes de datos

### 1.2.1 Matrices origen-destino generados por la red de telefonía móvil

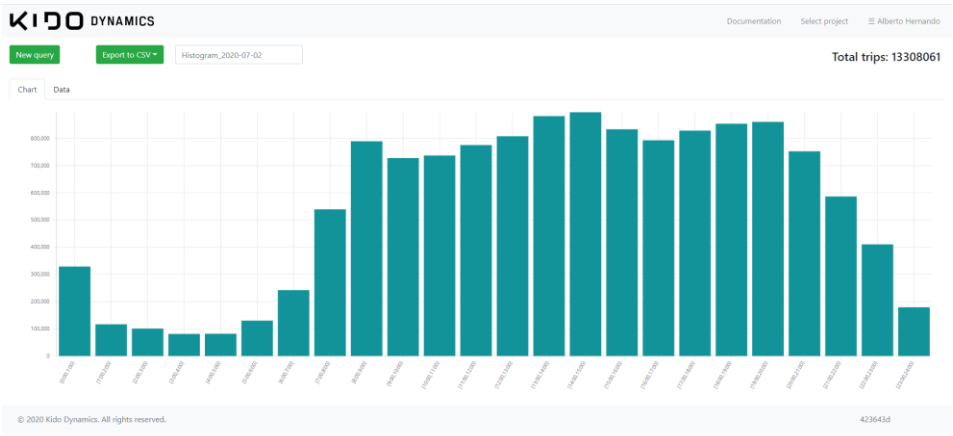
Los CDRs (Call Detail Records)[1] y los mensajes de señalización de la red móvil son registros que contienen datos relacionados con llamadas telefónicas, mensajes SMS, conexiones a Internet e información generada por la propia red para asegurar su control y la calidad de servicio. Desde el punto de vista de estudios de movilidad, cada registro proporciona localización espacio-temporal del dispositivo al relacionar su posición con una zona de cobertura (o "celda") asociada a una antena.

Estos datos, tras someterlos a un delicado procesamiento de Big Data (ver Anexo 1 en este documento) dan información muy precisa de los patrones de movilidad de días específicos, para todas las horas del día. La movilidad de una ciudad como Madrid queda descrita mediante matrices de origen-destino a un detalle muy fino (ver imagen), del que se pueden extraer series temporales para todo elemento origen-destino.

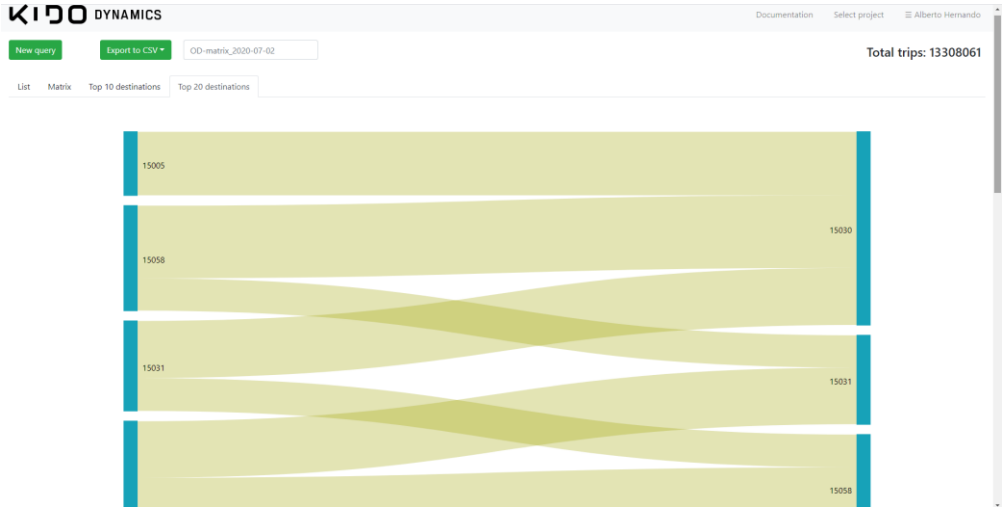



*Detalle de la grilla utilizada para describir la evolución en los patrones de movilidad en la ciudad de Madrid (www.kidodynamics.com/old/map.html).*

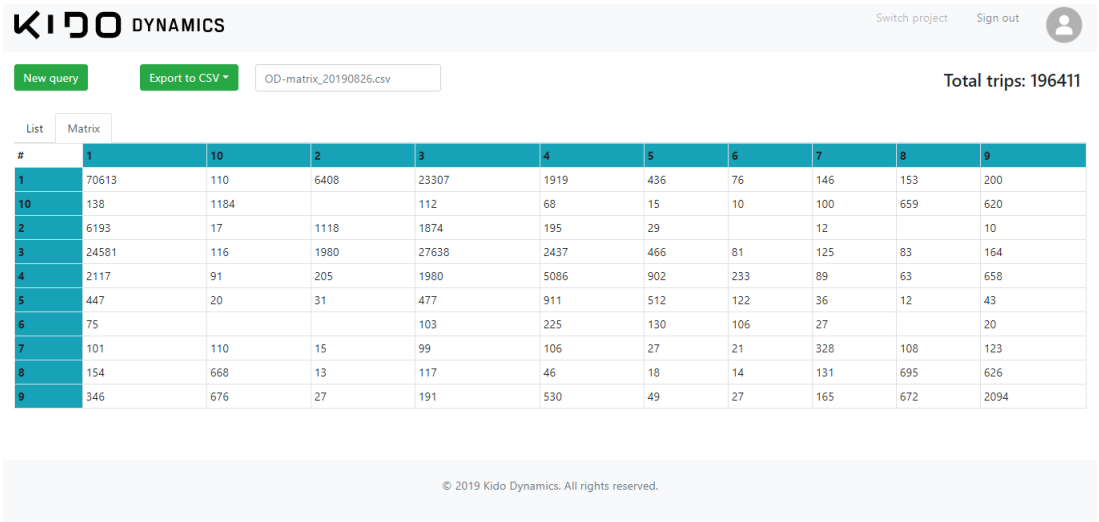
[1] Registros generados por las redes móviles como base de la tarificación a los usuarios.




*Serie temporal a lo largo de un día del número del total de viajes en una celda en particular (odmatrix.app).*




*Representación del número de viajes entre distintas celdas (origen a la izquierda, destino a la derecha).*




| # | 1 | 10 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 70613 | 110 | 6408 | 23307 | 1919 | 436 | 76 | 146 | 153 | 200 |
| 10 | 138 | 1184 | | 112 | 68 | 15 | 10 | 100 | 659 | 620 |
| 2 | 6193 | 17 | 1118 | 1874 | 195 | 29 | | 12 | | 10 |
| 3 | 24581 | 116 | 1980 | 27638 | 2437 | 466 | 81 | 125 | 83 | 164 |
| 4 | 2117 | 91 | 205 | 1980 | 5086 | 902 | 233 | 89 | 63 | 658 |
| 5 | 447 | 20 | 31 | 477 | 911 | 512 | 122 | 36 | 12 | 43 |
| 6 | 75 | | | 103 | 225 | 130 | 106 | 27 | | 20 |
| 7 | 101 | 110 | 15 | 99 | 106 | 27 | 21 | 328 | 108 | 123 |
| 8 | 154 | 668 | 13 | 117 | 46 | 18 | 14 | 131 | 695 | 626 |
| 9 | 346 | 676 | 27 | 191 | 530 | 49 | 27 | 165 | 672 | 2094 |

*Ejemplo de matriz origen-destino, con el número del total de viajes para cada par.*

Los datos de movilidad generados de esta forma proporcionan una visión completa de los movimientos de la ciudad, con valores absolutos en el número de personas que se han desplazado

entre los distintos puntos de la ciudad. Sin embargo, estas matrices no proporcionan por sí mismas información sobre el uso de transporte público o privado, ni del modo de transporte en general de forma directa. Es por tanto necesario cruzar con fuentes exteriores de datos para inferir el uso de la red de autobuses.

## 1.3 Datos del servicio de autobuses de la EMT

Para inferir la proporción de viajes que utilizan como modo de transporte la red de autobuses públicos e inferir la ocupación por bus, necesitamos por parte de la EMT:

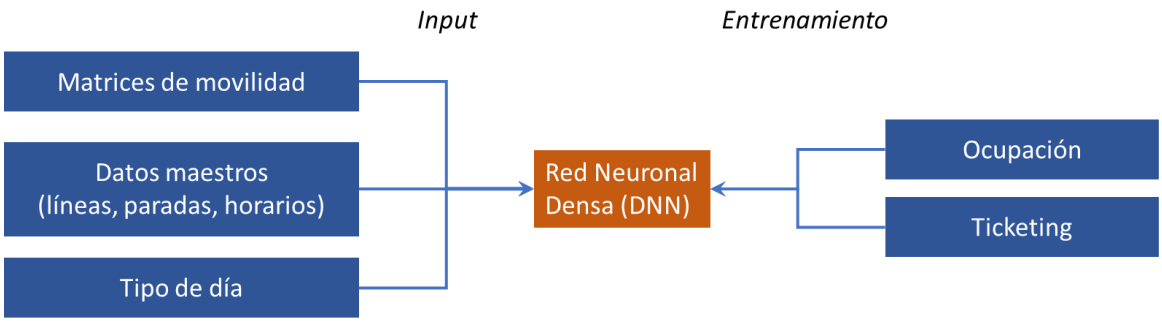
- Datos maestros: líneas, calendario de servicio, paradas, capacidad de los autobuses y modelos.
- Ticketing: Datos de los viajeros que han subido al autobús de los últimos tres años, desagregados por fecha, línea, autobús, viaje, sentido del recorrido, instante y parada.
- Ocupación: Datos de los dispositivos de conteo en el caso de aquellos autobuses en los que se disponga de esta información de manera fiable y contrastada.

Los datos maestros se utilizarán como base e input del modelo predictivo, mientras que el ticketing y la ocupación se usarán en el entrenamiento.

## 1.4 Procesos de entrenamiento, predicción e inferencia

### 1.4.1 Entrenamiento

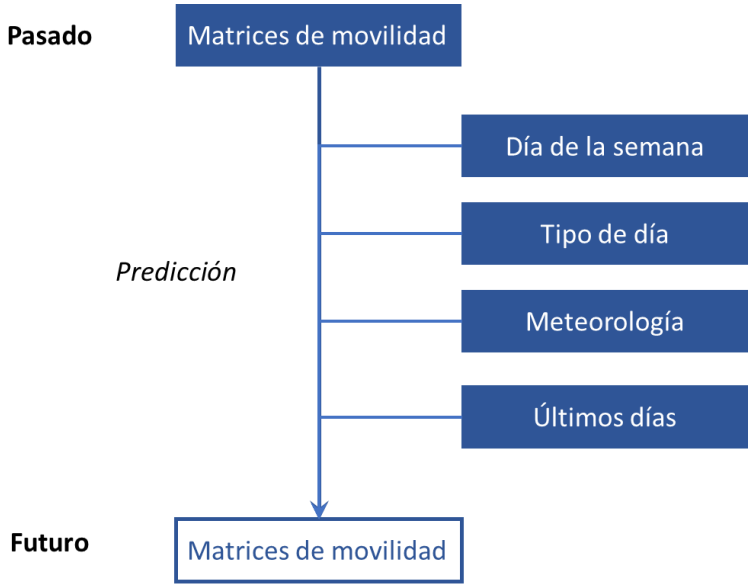
En esta fase se entrenará una red neuronal densa con el histórico de datos cruzados de movilidad y de ocupación.




### 1.4.2 Predicción

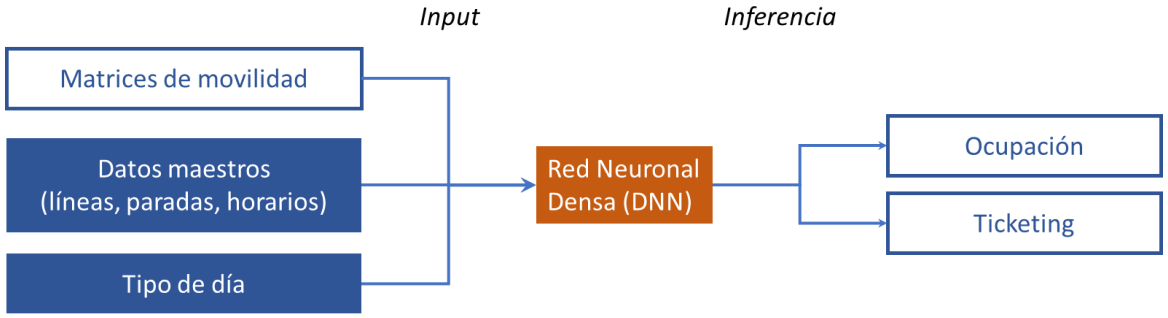
La predicción de las matrices de movilidad se realizará teniendo en cuenta la evolución de los días anteriores junto con los parámetros de tipo de día.

En el caso de eventos o movilidad excepcional, las matrices de movilidad se podrán modificar acordemente mediante modelos o estimaciones exteriores.




### 1.4.3 Inferencia

La inferencia se realizará a través de la red neuronal entrenada utilizando como input la predicción de la matriz de movilidad junto con la previsión de líneas, paradas y horarios programados para ese día.




## 1.5 Recursos técnicos

Todo el proceso se llevará a cabo con los recursos propios de software y hardware de Kido Dynamics. Nuestra arquitectura se describe en el Anexo 2 de este documento.

# 2 Descripción del análisis de metadatos móviles

## 2.1 Datos suministrados por el operador

### 2.1.1 Ficheros de comunicaciones generados por la Red Móvil

Los CDRs (Call Detail Records)[2] y los mensajes de señalización de la red móvil, son registros de que contienen datos relacionados con llamadas telefónicas, mensajes SMS, conexiones a Internet e información generada por la propia red para asegurar su control y la calidad de servicio. Desde el punto de vista de estudios de movilidad, cada registro proporciona información espacio-temporal del dispositivo relacionando su posición con una zona de cobertura asociada a una antena denominada celda.

En las planificaciones celulares de los operadores, el área de servicio de una celda suele superponerse parcialmente con el de una o más celdas vecinas. Este solapamiento tiene como objetivo evitar que haya zonas que se queden sin cobertura radioeléctrica.

Los registros de comunicaciones móviles generados por los diferentes elementos de la red móvil del operador (MSC, SMSC, MMSC, SGSN/GGSN) serán consolidados por sus sistemas de gestión y empaquetados en ficheros específicos por tipo de comunicación realizada. Estos ficheros incluyen todos los registros generados para los clientes del operador y clientes roamers que hacen uso de la red móvil como red visitante.

De cara a su procesamiento, estos ficheros se suelen identificar con su tipo de registro, el sistema que lo ha generado y el ámbito temporal al que corresponde. Desde una perspectiva general, los campos que contiene cada registro (señalización y CDR) de interés para las aplicaciones Big Data aplicadas a los estudios de movilidad son:

- Identidad del usuario: identificador anonimizado correspondiente a cada usuario.
- Tipo cliente: permite la identificación del usuario como cliente del operador responsable de la infraestructura o como usuario en tránsito (roamer).
- Fecha: Fecha de inicio de la comunicación.
- Hora: Hora de inicio y final de la comunicación
- Celda inicio: celda donde se localiza el móvil en el comienzo del registro.
- Celda final: celda donde se localiza el móvil al final del registro.

Algunos datos estadísticos cuantitativos

| CDRs | Señalización |
|---|---|
| • 14M de dispositivos vistos por mes<br>• 11M de dispositivos vistos por día<br>• 65% de dispositivos son vistos cada día del mes | • 12M dispositivos vistos por mes<br>• 11M de dispositivos vistos por día.<br>• 62% de dispositivos vistos cada día del mes. |

[2] Registros generados por las redes móviles como base de la tarificación a los usuarios.

| | |
|---|---|
| • 18% de dispositivos son vostos durante las 24h del día.<br>• 1.3B de eventos por día.<br>• 66% de dispositivos tienen menos de 100 eventos por día.<br>• 0.6% de dispositivos tiene más de 1000 eventos por día. | • 30% de dispositivos son vistos durante las 24 horas del día.<br>• 2.2B de eventos por día<br>• 46% de dispositivos tienen menos de 100 eventos por día.<br>• 1,5% tiene más de 1000 eventos por día. |

### 2.1.2 Ficheros de datos adicionales del operador a incorporar en los estudios de movilidad

Adicionalmente se utilizarán los siguientes ficheros:

- Fichero con información de la geolocalización de todas las estaciones base del operador correspondientes a las diferentes tecnologías desplegadas: 2G, 3G, 4G y 5G.
- Ficheros con información de clientes que será útil para incorporar la capacidad de segmentación solicitada y como referencia para la realización del "censo propio" en el proceso de expansión de la muestra que se define a continuación. Se incorporará la siguiente información:
    - Rango de edad
    - Género
    - Nacionalidad:
        - Españoles (identificando CP/provincia origen)
        - Extranjeros viviendo en España
        - Extranjeros de viaje por España (identificando cada país origen a partir de su operador)

## 2.2 Actividades específicas para el análisis de metadatos móviles

### 2.2.1 Filtrado, estructuración de datos y zonificación

Inicialmente, se realiza un análisis preliminar para asegurar la coherencia, calidad y anonimidad del flujo de registros generado por el operador móvil. Se filtran entradas erróneas y se compara la actividad de las comunicaciones y el comportamiento general en movilidad detectado por los registros del operador, con los patrones esperables y así asegurar su validez. Se incorporan los siguientes filtros:

1. Filtro de errores de transcripción. Se eliminan entradas erróneas o incompletas de los datos originales. Típicamente son antenas que no pueden localizarse, eventos con algún error de transcripción o campos vacíos. No son comunes, pero es necesario filtrarlos.
2. Filtro de masa crítica por usuario. El 25% de dispositivos con menor número de eventos se elimina. Son típicamente dispositivos con de 1 a 10 eventos por día, insuficientes para hacer un análisis de movilidad adecuado. Pese al filtro, el tamaño de la muestra sigue siendo de 12 a 13 millones de dispositivos.
3. Filtro de relevancia de la muestra. Por cada usuario, se filtran eventos redundantes o irrelevantes para la movilidad. Esto incluye secuencias de eventos en la misma antena o los

cambios de conexión a dos, tres y hasta cuatro antenas con solapamiento de cobertura pese a que el dispositivo no se haya desplazado.

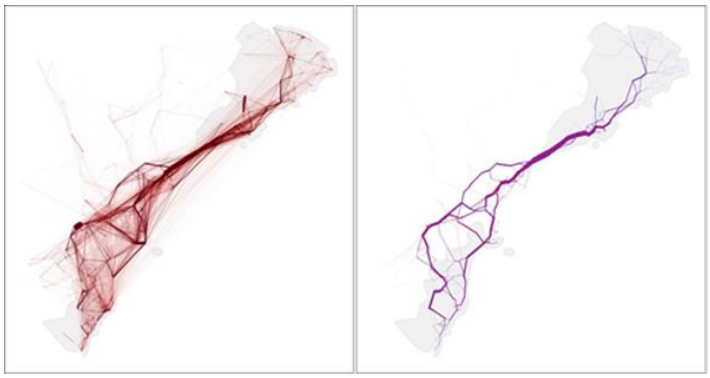
4. Filtro de acción. Se analiza la compatibilidad dinámica de la secuencia de eventos de cada usuario, filtrando aquellos que muestren un movimiento incompatible con la física en tiempo y espacio. Se da con dispositivos que comparten la misma identificación o si hay antenas mal localizadas. Dependiendo del origen de la incompatibilidad, se elimina el dispositivo o la antena.

Durante este proceso, también se filtran los eventos correspondientes a comunicaciones correspondientes a servicios M2M (machine to machine). En un día, de media pueden concurrir hasta 1.400.000 dispositivos de este tipo conectados a la red (casi un 10% del total) que no corresponden a usuarios reales, sino IoT, navegadores en vehículos, routers, etc. que también incorporan movilidad y podrían afectar a los resultados.

### 2.2.2 Reconstrucción de huecos

Las comunicaciones registradas por la red móvil no están asociadas a la posición GPS (sino a las posiciones de las antenas de la red), ni permiten obtener una trazabilidad de los desplazamientos continua en el tiempo. Las series temporales de eventos por dispositivo muestran saltos entre antenas en ocasiones de hasta varias horas y hasta varios kilómetros.

Para disponer de una visión completa de la movilidad se ha desarrollado un procedimiento de reconstrucción basado en los resultados de investigaciones realizadas en el campo de la sociofísica. La hipótesis principal es que la actividad demográfica humana responde al principio de máxima entropía. Aplicado en estudios de movilidad, la distribución de probabilidades más probable para la trayectoria de un usuario es aquella que maximiza la entropía en función de los datos de movilidad generados por el operador.



*Izquierda: el agregado de secuencias tras el filtro. Derecha: el agregado tras la reconstrucción.*

### 2.2.3 Agregación de usuarios

Una vez compleada la secuencia temporal de eventos por usuario, se agrupan sub-secuencias en estancias, microtrayectos, trayectos, y macrotrayectos en función de sus caracteríostocas cinemáticas. Las estancias y microtrayectos se utilizan en análisis de visitas, ocupación y aforos, mientras que los microtrayectos, trayectos y macrotrayectos se utilizan en análisis de tráfico.
A continuación se recogen las definiciones de cada uno de los conceptoes implicados:

1. Trayecto. Una secuencia de eventos se agrega dentro del mismo trayecto si el dispositivo ha recorrido al menos 5 km en el intervalo de la última hora. Este filtro asegura que dicho desplazamiento ha sido realizado con un vehículo a motor.

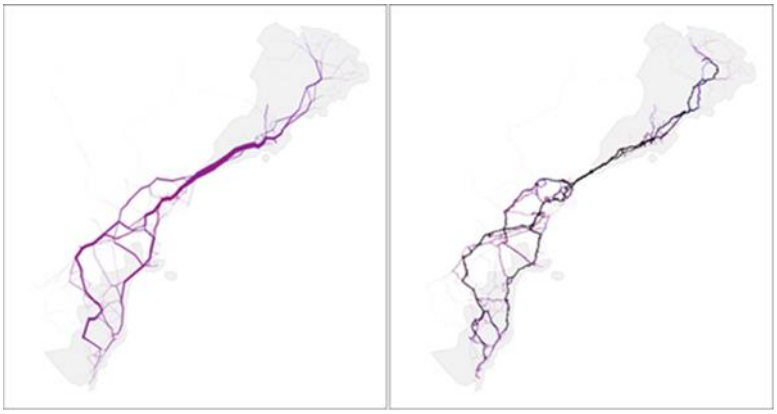
2. Microtrayecto. Una secuencia de eventos se agrega como microtrayecto si se ha producido un desplazamiento, pero inferior a la condición de trayecto del punto anterior.
3. Macrotrayecto o Viaje. Un macrotrayecto o viaje es una secuencia de trayectos y microtrayectos que conectan un origen con el destino final pese a intercalar momentos sin movilidad. La duración máxima de estos periodos intermedios se define por defecto tal que sea menor que la duración de las etapas que componen el viaje (el usuario podrá redefinir libremente esta duración para ajustarla a sus preferencias). Esta definición permite englobar viajes de larga distancia que hayan tenido paradas o viajes multimodales con intercambiadores.

Dentro del sistema, la duración de la estancia constituye un parámetro libre, de tal manera que utilizando una duración infinita de la estancia, se obtienen únicamente macrotrayectos, mientras que distintos valores de duración de las estancias permiten separar los macrotrayectos en trayectos y microtrayectos según las necesidades o criterios establecidos por el usuario.

### 2.2.4 Simulación de trayectorias en la red vial real

Finalmente, se implementa una nueva capa tecnológica que permite saltar de la localización espacial basada en antenas a una basada en el grafo de carreteras de la región estudiada, permitiendo un detalle sin precedentes de los patrones de movilidad. Para ello se generan millones de agentes que simulan la actividad observada en los eventos registrados por el operador móvil, para filtrar aquellos que maximizan la verosimilitud de sus acciones en el espacio y en el tiempo.

Tras la reconstrucción de la serie temporal por usuario, proyectamos el movimiento de la infraestructura de antenas a la infraestructura de carreteras según el camino más probable compatible cinemáticamente, en tiempo y distancia.



*Izquierda: el agregado tras la reconstrucción. Derecha: el agregado proyectado en la red de carreteras.*

*Ejemplo de trayectoria individual simulada con agentes virtuales (conjunto de trayectorias de la*

*izquierda) proyectada de la red de antenas (polígonos grises de la izquierda) al mapa de carreteras (derecha).*

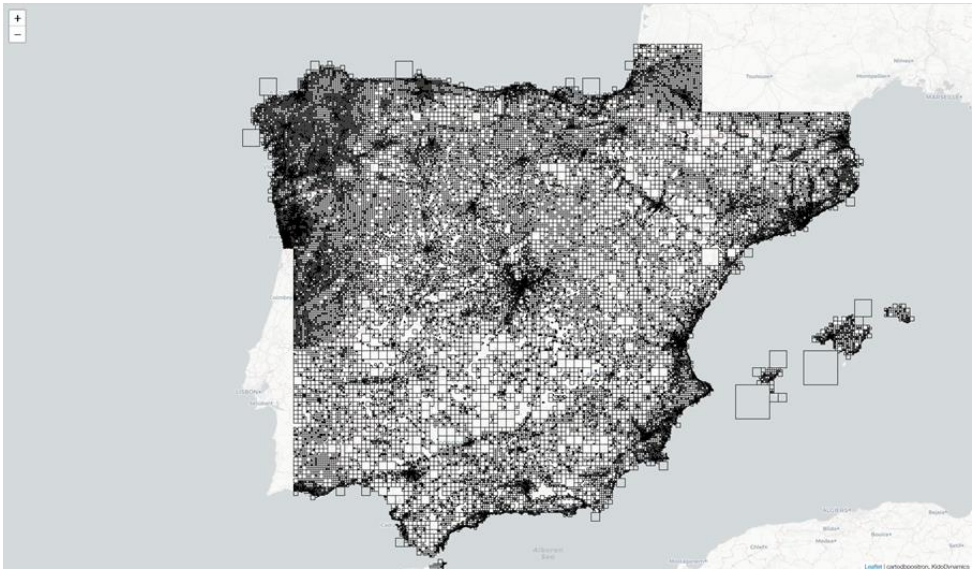
## 2.3 Garantía de privacidad del proceso

Tanto durante el proceso de análisis de los datos, como cuando los resultados se entregan al destinatario, se aseguran dos principios básicos: la incapacidad para identificar a una persona física, bien a partir de sus datos de identificación como a partir de los de ubicación. Las medidas para garantizar la privacidad son las siguientes:

1. Seudonimización. Los datos facilitados por el operador están previamente anonimizados mediante hash, sin acceso posible a ningún dato personal como número de teléfono, dirección, o nombre.
2. Privacidad diferencial I: anonimización individual completa. Al proyectar la secuencia de eventos a la infraestructura de carreteras, los agentes virtuales representando el movimiento de los dispositivos se desplazan siguiendo una componente aleatoria en su movimiento lo que hace imposible localizar puntos exactos como la vivienda o el lugar de trabajo de forma unívoca, haciendo imposible el cruce de datos a nivel individual.
3. Agregación. La contribución de cada dispositivo anonimizado es agregado por segmentos en bloques de un tamaño mínimo de 10 dispositivos. En caso de no agregar estos 10 dispositivos, el agregado es eliminado.
4. Privacidad diferencial II: anonimización de la agregación. Como medida adicional de seguridad, añadimos una capa de aleatoriedad a los resultados de la agregación. En este proceso añadimos a cada elemento un número aleatorio del orden de 20 dispositivos (uniforme entre -10 y +10).

## 2.4 Identificación de zonas de estudio

Para la definición de zonificaciones asociadas a proyectos de movilidad, se ha desarrollado un modelo de distribución territorial en coherencia con la demografía, la densidad de nodos de comunicación reales y la estructura de la red móvil. Para todo el territorio nacional, está constituido por aproximadamente 250.000 divisiones, lo que permite aumentar drásticamente el nivel de desagregación de las zonas de estudio de la movilidad en un ámbito geográfico determinado. También incorpora un avance respecto a la rigidez que supone tener que adaptarse a la estructura específica de áreas de cobertura de la red móvil del operador.



*Este modelo, que ha sido utilizado para definir las zonas núcleo y periféricas de las ciudades bajo estudio, permitiría aumentar la precisión en el detalle del análisis en áreas geográficas más específicas.*

## 2.5 Elevación de la muestra a población total

Consiste en la extrapolación de los resultados de la muestra de dispositivos que se ha utilizado para el análisis, al total de la población. Se realiza siguiendo la siguiente secuencia de acciones:

1. Asignación de residencia. Utilizando únicamente los eventos asociados a estancias, se asigna a cada dispositivo el lugar más probable de residencia utilizando como referencia los polígonos de sección censal según la definición del Instituto Nacional de Estadística (INE). La sección censal está diseñada de tal manera que asegura un nivel de agregación adecuado en todo el territorio español que garantiza la privacidad.
2. Creación del censo propio. Se agregan los resultados anteriores en una única lista para cada sección censal. Este censo se cruza con los datos demográficos de la lista de clientes del operados, asignando una pirámide de población propia a cada sección censal.
3. Cruce de censos. Para evitar un ajuste duro con el censo y el padrón (que se realiza anualmente) y perder fluctuaciones naturales de población flotante y otras dinámicas demográficas relevantes, se aplica un proceso estocástico que reproduzca la población a diferentes escalas de agregación. Esto se lleva a cabo aplicando a cada sección censal diversos radios de población (50.00, 10.000, 50.000, 100.000 y 500.000 habitantes), se cruza con los datos de cuota de mercado del operador, y se calcula para cada sección censal, por género y franja edad, el factor de escala necesario para equiparar la media de población de los distintos radios. El método permite compensar las fluctuaciones locales de la cuota de mercado a la vez que permite reproducir los cambios demográficos que se dan lugar en escalas de tiempo menores a un año.
4. Aplicación de pesos estadísticos. Una vez obtenidos los factores de escala por cada segmento demográfico, se aplica a los agregados en las mismas proporciones para elevar los resultados a total de habitantes en lugar de dispositivos móviles.

Los dispositivos en roaming-in (típicamente visitantes internaciones en territorio español) carecen de datos demográficos y sólo se dispone de la nacionalidad de su operador. En este caso la elevación se obtiene por cada nacionalidad utilizando de referencia los datos del FRONTUR y por cada operador según los acuerdos internacionales de roaming entre operadores.

Con esta metodología pueden obtenerse resultados de movilidad eliminando los sesgos derivados de la duplicidad de terminales móviles en el mismo usuario o de la edad del usuario del terminal móvil frente a la edad del titular de la línea contratada, así como permite la detección de población flotante que pueda residir en una ciudad sin que necesariamente esté registrada en su lista del censo.

Para proteger la privacidad de los usuarios en la eliminación de estos sesgos, cono se ha indicado previamente, en ningún momento se recurre a un seguimiento de los comportamientos en movilidad.

# 3 Recursos técnicos Hardware y Software

## 3.1 Infraestructura en la nube desplegada por KIDO DYNAMICS

KIDO DYNAMICS para la ejecución del proyecto contará con la infraestructura tecnológica en la nube AMAZON WEB SERVICE [3](ya desplegada actualmente).

En este momento, KIDO DYNAMICS dispone de acceso con continuidad a los metadatos de movilidad del Operador Móvil seleccionado para este proyecto (Orange), tanto CDRs como metadatos de señalización.

Actualmente, están disponibles en la plataforma de servicios de KIDO DYNAMICS los metadatos de la red móvil correspondientes a CDRs de todos los usuarios móviles del operador seleccionado desde el 1 de enero de 2019 y también datos de señalización desde el 1 de enero de 2020. También se dispone de las herramientas seguras apropiadas para la gestión de datos en la nube privada en Amazon Web Services.

Todos los recursos necesarios para la realización del proyecto serán contratados de forma permanente por KIDO DYNAMICS al proveedor tecnológico AMAZON en el momento de la adjudicación

## 3.2 Arquitectura basada en servicios AWS

La nube de KIDO DYNAMICS consta de dos partes:

- Un entorno gestionado y auditado por el operador, donde el equipo técnico propio del operador copia los CDR sin procesar y los datos de señalización en un depósito común de solo lectura.
- Un segundo entorno seguro administrado y propiedad de KIDO DYNAMICS donde los datos se vuelven a codificar para su seudonimización (asegurando que se mantienen separados de las series de datos de cualquier otro usuario) y donde tiene lugar la agregación posterior. Los resultados agregados se generan en este entorno para alimentar nuestras aplicaciones y la API REST definida para los servicios de KIDO DYNMICS en transporte, turismo, venta minorista o marketing.

El sistema de KIDO DYNAMICS permite definir dos tipos de usuarios.:

- Kido Developer: acceso completo en ambos entornos. Permisos para los servicios de AWS para implementar máquinas virtuales, almacenamiento y entrega y administración de datos.
- Usuario final: acceso solo a través de las aplicaciones en línea SaaS de Kido. Todos los inicios de sesión están asegurados y todas las acciones realizadas en la cuenta de AWS de Kido se registrarán con AWS CloudTrail para futuras auditorías.

## 3.3 Tecnologías, herramientas software y servicios

La tecnologías, herramientas software específicas y servicios más relevantes desplegados en la infraestructura tecnológica de KIDO DYNAMICS son:

### Computación en la nube

- Athena[4]
- EMR[5]

### Computación paralela

[3] https://aws.amazon.com

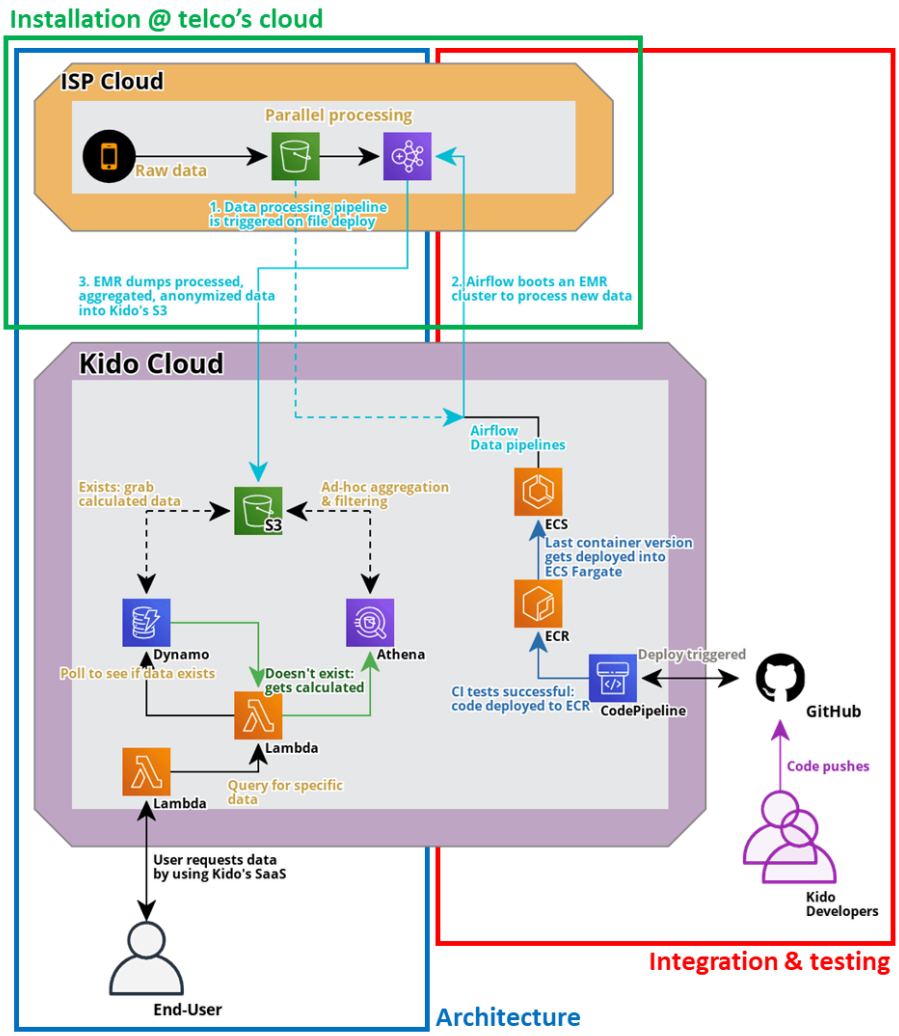
[4] https://boto3.amazonaws.com/v1/documentation/api/latest/reference/services/athena.html

[5] https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-what-is-emr.html

- PrestoDB[6]
- Spark[7]

Auditoría en la nube

- CloudTrail[8]




*Arquitectura en la nube de KIDO DYNAMICS basada en servicios AWS*

[6] https://prestodb.io/docs/current/index.html

[7] https://spark.apache.org/

[8] https://docs.aws.amazon.com/awscloudtrail/latest/userguide/cloudtrail-user-guide.html