MEJORA DE LA CARACTERIZACIÓN DE LOS SERVICIOS DE UN SISTEMA DE TRANSPORTE PÚBLICO COMBINANDO DATOS DE LOS SUBSISTEMAS DE LOCALIZACIÓN, BILLETAJE Y PLANIFICACIÓN

Juan Benavente Ponce
Universidad de Cantabria
Borja Alonso Oreña
Universidad de Cantabria
José Luis Moura Berodia
Universidad de Cantabria
Andrés Rodríguez Gutiérrez
Universidad de Cantabria

RESUMEN

Este trabajo propone una metodología flexible para mejorar la caracterización de cada uno de los servicios ofrecidos por un sistema de transporte público, utilizando eventos de llegada de autobuses a nivel de parada y de embarque de pasajeros procedentes de sus sistemas automatizados de localización de vehículos y de cobro de billetes, así como la información del subsistema de planificación si está disponible. Se aplican diferentes técnicas para estructurar, corregir y completar los datos disponibles, con el objetivo de minimizar las distorsiones que aparecen debido a la naturaleza de estas fuentes.

El procedimiento descrito en este trabajo pretende ser adecuado en situaciones con diferente disponibilidad, integridad y fiabilidad de la información. En particular, los inicios de los servicios programados pueden ser conocidos o no, incluyendo opcionalmente qué vehículo había sido asignado inicialmente a la tarea. La información procedente de los diferentes subsistemas que integran un sistema de ayuda a la explotación se combina para crear una caracterización mejor y más útil de los servicios que han tenido lugar en un sistema de transporte.

Los datos del ejemplo de aplicación provienen de los eventos del sistema de ayuda a la explotación en la ciudad de Santander (España) durante un año. Los resultados se discuten con capturas de una herramienta de visualización web interactiva que se ha desarrollado para este trabajo.

1. ESTADO DEL ARTE

Los sistemas de transporte público inteligentes (*IPTS*) están construidos sobre varios subsistemas que se encargan de diferentes tareas, como proporcionar información a los viajeros o trabajadores, gestionar incidencias y eventos especiales, localizar los vehículos, programar, emitir billetes, contar pasajeros, información geográfica, nóminas, mantenimiento, meteorología, satisfacción del cliente o comunicaciones. Si se implantan con éxito, aumentan la calidad del servicio, disminuyen los costes de explotación, mejoran el proceso de toma de decisiones y facilitan la gestión de la flota [de Pablos Heredero et al., 2012].

Dejando a un lado los casos en los que no se dispone de horarios fijos (por ejemplo, las rutas de autobús en Jinan (China), con una gran incertidumbre en los tiempos de viaje, múltiples agencias y un horario de salida que cambia según la demanda observada in situ, donde un estudio empleó redes neuronales artificiales para mejorar la estimación de la llegada de los autobuses en tiempo real basándose en la información del sistema de billetaje automático (AFC) y de la ubicación histórica de los vehículos [Lin et al., 2013]), los datos que describen los servicios de tránsito (es decir, las rutas, sus horarios y la ubicación de las paradas) se publican con antelación. La herramienta más extendida para hacerlo es el componente estático de la especificación general de datos de tránsito (GTFS) [Google]. Sin embargo, aunque se ha propuesto una extensión [MobilityData], este formato aún no puede representar algunos cambios en tiempo real, como la definición de viajes adicionales. Además, es posible que las agencias de transporte no mantengan una recopilación de estos archivos a lo largo del tiempo, aunque en algunos casos pueden obtenerse de terceros (por ejemplo, OVapi [OVapi B.V], Transitland [Mapzen Foundation e Interline Technologies LLC], u OpenMobilityData [MobilityData IO,]).

Además de otras aplicaciones como la identificación de irregularidades en las cabeceras [Gokasar y Cetinel, 2019], la implementación de estrategias de prioridad vehicular más inteligentes [Hounsell y Shrestha, 2012], y la gestión de flotas y operaciones [Saghaei, 2016]; el satélite de navegación global por satélite (GNSS) se utiliza para estimar inicialmente y finalmente identificar la llegada del vehículo a cada punto de interés. Normalmente, ofrece una precisión de 5 m a cielo abierto, aunque hay varios factores que pueden empeorar su precisión [van Diggelen, Frank, Enge, 2015].

Por otro lado, los sistemas AFC tienen como principal objetivo mejorar el proceso de recaudación, pero también proporcionan datos valiosos, especialmente cuando se enriquecen con las posibilidades de seguimiento y caracterización de usuarios de la tecnología de tarjetas inteligentes (SC).

Cuatro aspectos clave caracterizan la información de los IPTS [Furth et al., 2003]: detalle espacial y temporal, cobertura (todos los eventos o sólo los excepcionales), representatividad (penetración de la flota y tasa de recuperación de datos) y calidad.

Los registros a nivel de parada, que pueden almacenarse en el IPTS con un coste incremental bastante bajo, han permitido estimar mejor los indicadores de rendimiento y las métricas de uso anteriormente utilizados (por ejemplo, los tiempos de viaje) [Trépanier et al., 2009], y también evaluar atributos antes casi imposibles de cuantificar debido a la escasez de datos, como los relacionados con la fiabilidad del servicio [Wilson et al., 2009]. Sin embargo, pueden requerir un esfuerzo importante para obtener conclusiones significativas [Ma et al., 2014]. Además, se necesitan herramientas de visualización adecuadas para poder comprender la gran cantidad de resultados que se pueden generar. Bertini y El-Geneidy, 2003].

En la literatura existente se proponen varias distribuciones de probabilidad para caracterizar la variabilidad de los tiempos de viaje de los enlaces [Dai et al., 2019], como la log-normal desplazada, la log-normal, la normal [Qu et al., 2014], la gamma, la de Weibull, la de Burr tipo XII [Taylor y Susilawati, 2012], la de valor extremo generalizado [Chepuri et al., 2018], etc. Numerosos estudios [Harsha et al., 2020, Dai et al., 2019, Li et al., 2017, Srinivasan et al., 2014] eligen la primera, que muestra una densidad de probabilidad de cero cuando el valor de la variable aleatoria cae por debajo de un umbral (que sería el tiempo de viaje del enlace de flujo libre) y puede ajustarse adecuadamente a los datos asimétricos y con sesgo positivo; y que para muchos enlaces es la función que más probablemente describe cómo se distribuyen los tiempos de viaje. Un estudio de 2017 realizado con datos del sistema de posicionamiento global (GPS) de los taxis durante las horas punta de la mañana de 5 días laborables en Wuhan (China) [Chen et al., 2017] descubrió que los tiempos de viaje de los enlaces pueden estar mejor representados por distribuciones logarítmicas normales, gamma o normales (en el 50 %, el 30 % y el 20 % de los enlaces analizados, respectivamente) y optó, para evitar cálculos intrincados desde el punto de vista informático, por asumir que los tiempos de viaje a lo largo de un trayecto pueden aproximarse mediante distribuciones normales.

En cuanto a los tiempos de permanencia, la mayoría de los trabajos sugieren que, debido a su naturaleza no negativa y a su posible asimetría, es probable que la distribución logarítmica normal sea la mejor alternativa (por ejemplo, un estudio de 18 meses de datos de una ruta de autobús en Nueva Jersey, Estados Unidos [Rajbhandari et al., 2003]; 6000 registros de un estudio de un día en Changzhou, China; o un análisis de datos de un mes de autobuses públicos en Jinan, China [Zhang et al., 2019]). Otras distribuciones posibles son la normal, utilizada por el software comercial de microsimulación de tráfico como AIMSUN [Aimsun, 2020] o VISSIM [PTV Group, 2020], y también elegida en algunos trabajos científicos (por ejemplo para caracterizar los datos de 1 día de una parada de autobús en la ciudad de Chennai, India [Koshy y Arasan, 2005]); Wakeby, que superó a la distribución log-normal

en un estudio con 3 meses de datos de 4 paradas en Auckland,Nueva Zelanda [RASHIDI y RANJITKAR, 2013]; o Erlang, propuesta en un estudio que analizó 435 registros de 12 paradas de autobús en Shanghái, China [Jiang y Yang, 2014].

Los subsistemas que contribuyen a un IPTS a menudo no captan adecuadamente información que sería útil para un análisis posterior, porque suelen tener otros objetivos: apoyar la planificación táctica y la respuesta de emergencia en el caso del AFC, y gestionar las concesiones para el AFC. En consecuencia, suelen surgir una serie de cuestiones relacionadas con problemas internos de cada conjunto de datos o con incoherencias entre ellos. Los que entran en el ámbito de este trabajo son [Luo et al., 2018]:

- Registros AFC erróneos, que pueden ser causados por fallos de funcionamiento, comportamiento atípico de los viajeros, desvíos de rutas de emergencia o mal manejo de los equipos por parte de los conductores y operadores [Trépanier et al., 2007].
- Entradas erróneas en el sistema de localización automática de los vehículos (*AVL*) debido a fallos del del mismo, operaciones incorrectas del controlador o problemas específicos en las terminales.
- Múltiples registros para el mismo evento AVL, posiblemente con diferentes atributos (marca de tiempo, identificación del vehículo o de la ruta).
- Eventos perdidos.
- Falta de información que vincule los viajes reales con el programa de servicios vigente en ese momento.
- Ausencia de información o información errónea para relacionar los viajes con los desplazamientos de los vehículos.

En algunos casos, estos problemas pueden ser tan graves que los investigadores han desarrollado metodologías que modelan las características del transporte público de forma indirecta, en lugar de utilizar una alternativa más inmediata, pero propensa a errores (por ejemplo, utilizando AVL en lugar de AFC o registros de contadores de pasajeros automatizados para estimar la demanda de transporte público [Moreira-Matías, 2016]).

Hay muchos ejemplos publicados de la aplicación combinada de múltiples sistemas de datos de recogida automatizada en los diferentes aspectos de la gestión y planificación del tránsito urbano. Entre los que utilizan datos de AVL y AFC, algunos ejemplos destacables son:

- Perfiles de carga espaciotemporales de los vehículos de tránsito urbano durante un mes en La Haya (Países Bajos), integrando completamente los registros GTFS como tercera fuente de datos con la información AVL y AFC [Luo et al., 2018].
- Procesamiento fuera de línea de los sistemas de seguimiento automatizado de trenes y de cobro de billetes basados en tarjetas magnéticas de viaje en el área de la bahía de San Francisco (EE.UU.) [Buneman, 1984].

- Estimación de matrices origen-destino (OD) y modelos de elección de ruta para pasajeros de ferrocarril de la Autoridad de Tránsito de Chicago [Wilson et al., 2009].
- Modelización multimodal del propósito del viaje y estimación mejorada de la DO en Queensland (Australia) [Alsger y Eng, 2016].Matrices de OD de metro y autobús, perfiles de velocidad de los vehículos y calidad, indicadores de servicio, etc. para el sistema de transporte público Transantiago en Santiago de Chile [Gschwender et al., 2016]. "Entrevistas de autobús asistidas por el conductor": si los registros de SC están correctamente vinculados con la información de AVL, pueden funcionar como encuestas de preferencias reveladas [Chu et al., 2009].
- Seguimiento de los SC a lo largo del metro y el autobús para identificar el comportamiento de los transbordos en Shenzhen (China), haciendo uso de los registros del AFC del autobús que solo muestran la identificación de la tarjeta y el tiempo de barrido [Huang et al., 2019].

Sin embargo, no hemos encontrado en la literatura existente una metodología que se centre en mejorar la caracterización de los servicios que tienen lugar en un sistema de transporte público de autobuses combinando AFC, AVL, y arranques programados; con el objetivo de trabajar en torno a los problemas comunes de datos de IPTS y, lo suficientemente flexible como para ser aplicado en diferentes situaciones. Esperamos que sea útil para otros investigadores e ingenieros de transporte durante sus actividades; como la auditoría, el modelado del comportamiento de los usuarios o la estimación de los perfiles de carga de los vehículos.

2. METODOLOGÍA

Esta sección comienza especificando las fuentes y la estructura prevista de los datos de entrada. A continuación, se detallan los pasos de preprocesamiento que se aplican a la información AVL, AFC y de planificación; representando cada visita de un autobús a una parada como un único evento de cada fuente. A esto le sigue el análisis de los datos AVL como secuencias que se descomponen en fragmentos de sus respectivas líneas; y la implementación de modelos de distribución de tiempos de viaje de los enlaces y de tiempos de permanencia. A continuación, se lleva a cabo la caracterización inicial de los servicios realizados, que puede mejorarse si es necesario detectando cambios de id de los vehículos a mitad de servicio. A continuación, se vinculan los inicios programados para identificar los servicios planificados y los adicionales, distinguiendo en el primer caso si se utilizó el vehículo previsto o no; y también para mejorar la fidelidad de la recreación. A continuación, se asignan los eventos AFC a las visitas del autobús. Por último, se aceptarán aquellos servicios respaldados por suficiente información del IPTS. La figura 1 muestra un resumen general de todo el proceso.

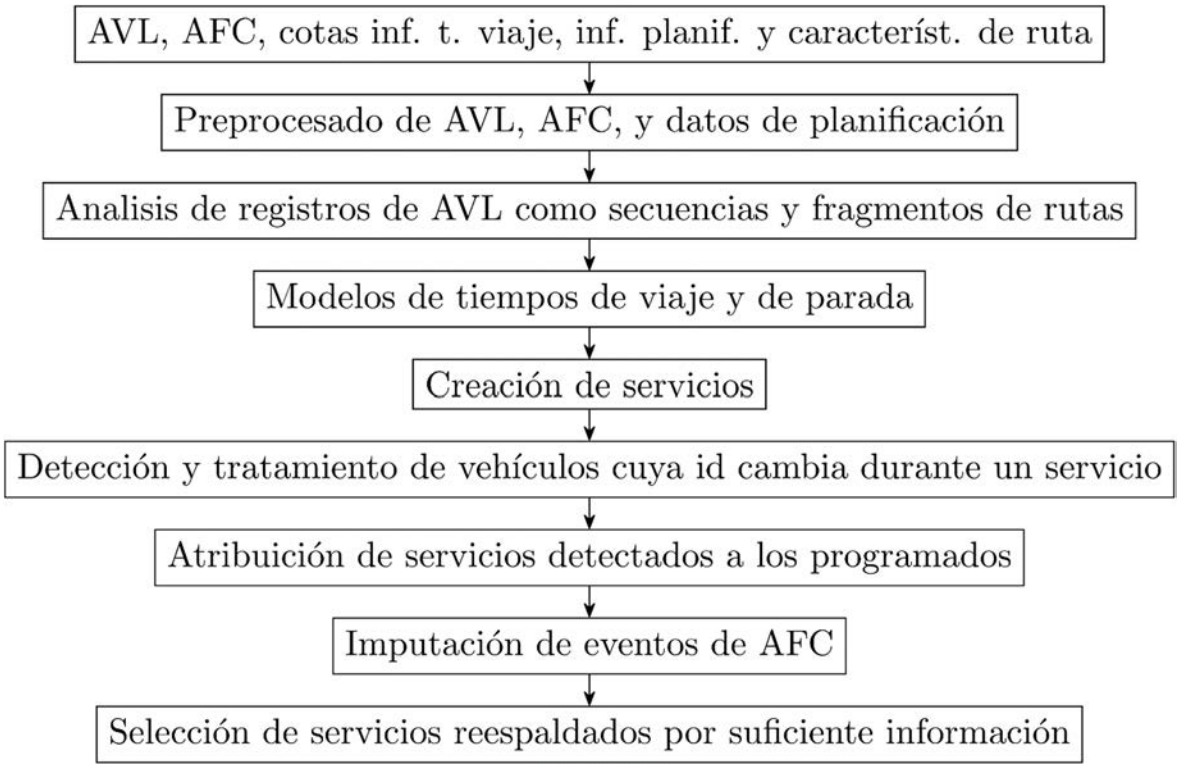


Figura 1: Esquema de la metodología

2.1 Datos de entrada

La Tabla 1 contiene un resumen de la información necesaria sobre las paradas de autobús, AFC, AVL, límites inferiores de los tiempos de viaje y el horario de inicio del servicio. Cabe señalar que los identificadores de las paradas de autobús, las rutas y los vehículos deben ser coherentes en todos los subsistemas. Las columnas llamadas “grupo” en AFC y AVL deben contener un identificador único para cada conjunto de valores presentes en su subsistema particular, que puede ayudar a diferenciar entre los recorridos de un vehículo.

En cuanto al horario, la metodología está diseñada para funcionar incluso cuando está incompleto, o para detectar recorridos de servicios no planificados. En esta sección se asumirá que las tres columnas con información temporal pueden estar disponibles en al menos parte del conjunto de datos.

Columna	*Tipo*	*Descripción*
id	entero	Id del bus en el IPTS.
localización	(real, real)	Coordenadas geográficas.
nombre	texto	Cómo es llamada por los usuarios.

(a) *paradas*

Columna	*Tipo*	*Descripción*
parada_0	entero	Id de la parada inicial.
parada_1	entero	Id de la parada final.
t_fluj_libr	tiempo	Cota inferior del t. necesario para este viaje.

(b) *cotas inferiores de los tiempos de viaje*

Columna	*Tipo*	*Descripción*
parada	entero	Id de la parada.
línea	entero	Id de la línea.
vehículo	entero	Id del vehículo.
t_validac	instante	Momento de validación.
grupo	entero	UID para cada conjunto de valores de otras cols. que varían en diferentes visitas del bus.

(c) *AFC en bruto*

Columna	*Tipo*	*Descripción*
parada	entero	Id de la parada.
línea	entero	Id de la línea.
vehículo	entero	Id de vehículo.
durac_parad	tiempo	Cuánto t. estuvo el bus en la parada.
llegada	instante	Momento de llegada a la parada.
grupo	entero	UID para cada conjunto de valores de otras cols. que diferencien distintas visitas de un bus.

(d) *AVL en bruto*

Columna	*Tipo*	*Descripción*
línea	entero	Id de línea.
vehículo	entero	Id de vehículo.
parada	entero	Id de parada.
comienzo_pl	instante	Instante planificado de inicio.
llegada_reg	instante	Detecc. llegada vehículo a parada inicial.
comienzo_reg	instante	Detecc. comienzo servicio.

(e) *planificación en bruto*

Columna	*Tipo*	*Descripción*
línea	entero	Id de línea.
term_dudos	lógico	Verdadero si AFC y AVL en cabec. poco fiables.
ini_srv_crrct	tiempo	T. medio entre 1^{er} evento AVL y detecc. inicio serv.
sep_max	tiempo	Cota superior entre el paso de dos servicios.
t_prds_max	tiempo	T. viaje máx. entre paradas consecutivas.
t_vuelt_min	tiempo	T. min. para dar una vuelta completa.

(f) *información de las líneas*

Tabla 1: Datos de partida de la metodología

2.2 Preprocesamiento

En esta sección se crean nuevas tablas para los conjuntos de datos AFC y AVL, sintetizando en una sola entrada la información que cada fuente bruta proporciona sobre una visita de autobús. Además, se utilizan límites inferiores de tiempos de viaje para filtrar los datos AVL poco fiables.

2.2.1 AFC

Se supone que no hay filas duplicadas en la información bruta de AFC, ya que debido a las repercusiones monetarias de los datos, la información de los billetes se gestiona de forma muy cuidadosa. Las operaciones de pago con tarjeta inteligente o manuales son atómicas: o se completan con éxito o no se producen.

Como se explicará en detalle en la sección 2.5, la información del AFC (que proporciona un dato por validación) se utiliza para hacer frente a las carencias de los datos del AVL (idealmente, un dato por visita al autobús). Así, el objetivo es clasificar como un único evento de embarque todas las validaciones que se producen cada vez que un autobús hace escala en una parada. El primer y último eventos de emisión de billetes de estos "grupos de embarque" pueden utilizarse como una aproximación de cuándo llegó y salió el autobús de la parada. Para crearlos, se lleva a cabo un proceso de tres partes:

2.2.1.1 Crear grupos de parada

Se analizan los registros en bruto de AFC correspondientes a cada autobús, buscando distinguir grupos de eventos consecutivos referidos a la misma visita a una parada.

id parada	*rango en conj.*	*rango en subconj.*	*variable clasiff.*	*grp. parad.*	*instante validación*	*sep. temp.*	*cumple lím. sup.*	*# AVL intermedio*	*cambio. grupo embarq.*	*grupo de embarq.*
					...					
A	1	1	0	A	03-24 12:31:23	<null>			1	987
B	2	1	1	B	03-24 12:33:56	<null>			1	988
C	3	1	2	C	03-24 12:36:14	<null>			1	989
C	4	2	2		03-24 12:36:16	00:00:02	✓	✓*	0	
D	5	1	4	D	03-24 12:37:24	<null>			1	990
E	6	1	5	E	03-24 12:39:44	<null>			1	991
E	7	2	5		03-24 12:45:22	00:05:38	✓	✓*	0	
F	8	1	7	F	03-24 12:47:37	<null>			1	992
G	9	1	8	G	03-24 13:48:51	<null>			1	993
G	10	2	8		03-24 13:50:59	00:02:08	✓	✓*	0	
G	11	3	8		03-24 13:53:04	00:02:05	✓	✓*	0	
G	12	4	8		03-24 14:11:11	00:18:07	✓	✗**	1	994
G	13	5	8		03-24 14:11:13	00:00:02	✓	✓*	0	
					...					
B	45	2	43	B	03-24 15:49:28	<null>			1	1004
C	46	3	43	C	03-24 15:51:33	<null>			1	1005
D	47	2	45	D	03-24 15:54:02	<null>			1	1006
D	47	2	45		03-24 15:54:02	00:00:00	✓	✓*	0	
D	48	3	45		03-24 23:05:49	08:11:47	✗		1	1007
E	49	2	47	E	03-24 23:09:09	<null>			1	1008
					...					

*: No se encontró una visita intermedia a otra parada en la información AVL
**: Los registros AVL revelan que este vehículo visitó la parada K a las 14:01:51

Tabla 2: Preprocesado de AFC en bruto. Vehículo, línea y UID de grupo permanecen constantes durante este ejemplo.

Este procedimiento se basa en el hecho de que, tal y como se representa en la tabla 2, para un conjunto (en este caso, las entradas del AFC en bruto vinculadas a un único vehículo) en el que se puede utilizar una relación ("ocurrió antes") para establecer una jerarquía sobre todo éste (columna "rango sobre el conjunto") y también sobre los diferentes subconjuntos definidos por una partición (entradas con los mismos valores de *vehículo*, *línea*, *grupo*, e *id de parada*) la diferencia entre el rango sobre el conjunto y sobre un subconjunto particular (la columna "variable de clasificación") proporciona un valor que diferencia los miembros de ese subconjunto que aparecen consecutivamente al ordenar todos los elementos del conjunto (el grupo de paradas, mostrado en la columna "grp. parad."). El significado del resto de las columnas de la tabla 2 y la coloración de las celdas se explicará a medida que se mencione a lo largo del resto de esta descripción del preprocesamiento del AFC. El proceso se explica como tres tareas consecutivas:

2.2.1.1.1 Clasificación por vehículo

Las entradas del AFC se agrupan por vehículo, y luego se obtiene su rango según una ordenación cronológica.

2.2.1.1.2 Clasificación por vehículo, ruta, grupo y parada de autobús

Las entradas AFC se clasifican por vehículo, línea, grupo y parada de autobús; y de nuevo se almacena su rango tras una ordenación cronológica. Estas cuatro columnas de la tabla de datos brutos de AFC permanecen constantes durante todas las validaciones de un evento de embarque concreto.

2.2.1.1.3 Crear grupos de parada

La diferencia de los dos rangos descritos anteriormente, la '*variable de clasificación*', es constante y única para cada grupo de filas que reportan los mismos valores de *vehículo*, *línea*, *grupo* y *parada*, y aparecen consecutivamente.

La columna '*grupo parada*' de la tabla 2 muestra el resultado de esta primera aproximación al objetivo de identificar los grupos de embarque; mostrando una sola letra para todas las entradas de avl en bruto consecutivas que forman parte del mismo grupo de parada. La coloración de las columnas '*id. de parada*', '*variable clasificación*' y '*grupo de parada*' ilustra el proceso de clasificación y su resultado. Por ejemplo, las filas correspondientes a las visitas a la parada D se reúnen en dos '*grupos de paradas*', diferenciados por los valores de la 'variable de clasificación' de 2 y 5.

Una forma de comprobar cómo ha funcionado esta primera tarea es estudiar la '*separación temporal*' (tabla 2) entre las entradas consecutivas del mismo 'grupo de paradas'. A medida que este desfase aumenta, es más probable que esta última entrada haya tenido lugar durante una visita diferente del autobús (sin filas intermedias debido a que no se registraron validaciones hasta que el autobús regresó). En la siguiente sección se estudia esta situación.

2.2.1.2 Dividir grupos de parada en grupos de embarque

La cuestión relativa a los desfases temporales excesivos entre algunas de las entradas que forman parte del mismo grupo de paradas se aborda con los siguientes supuestos:

- En algunas ciudades, no es raro que el conductor permita a los pasajeros esperar el inicio de un servicio dentro del autobús, especialmente si hace mal tiempo. Sin embargo, si la separación entre dos entradas consecutivas del mismo grupo de paradas es mayor que el recorrido máximo de la ruta, su grupo se dividirá entre ellas. Esto ocurre en la penúltima fila de la tabla 2: el tiempo transcurrido desde la validación anterior es extremadamente largo (representado con el símbolo '✓' en la columna '*cumple límite superior*'), por lo que se puede estar seguro de que está describiendo un evento de embarque diferente y el grupo se divide, como se ha representado con el cambio de color de naranja a rojo. Un valor adecuado para este parámetro dependerá de las particularidades del caso analizado.
- Para todos los demás pares de entradas consecutivas del mismo grupo de paradas, si la tabla *avl_sintetizado* (definida más adelante durante la descripción del preprocesamiento de los datos AVL) muestra que el autobús visitó otra parada entremedias, pertenecen a grupos de embarque diferentes. Un ejemplo de esta situación se encuentra en la fila con '*rango en conjunto*' = 12 de la tabla 2, donde el grupo de paradas de 5 entradas (G,8) del que forma parte está dividido en dos grupos de embarque (993 y 994); porque, como denota el símbolo '✗' de la columna '∄ *registro AVL intermedio*', entre su marca de tiempo (14:11:11) y la de la entrada anterior (13:53:04) una búsqueda a través de los datos AVL brutos (no representados) ha concluido que el autobús pasó por la parada K a las 14:01:51.

Estas premisas se utilizan para definir la columna '*cambio de grupo de embarque*' de la tabla 2), un valor que será igual a 1 si una fila es la primera de un grupo de embarque, y 0 en los demás casos. Si las filas se ordenan por vehículo, rango cronológico, parada de autobús y cambio de grupo de embarque (de forma descendente), aparecerán consecutivamente si forman parte del mismo grupo de embarque, con un valor de cambio de grupo de 1 para la primera entrada y de 0 para las demás hasta su final. El identificador de grupo de embarque de cada fila es la suma total de todos los cambios de grupo hasta llegar a ella.

Volviendo a la Tabla 2, el contenido y los colores de las celdas de las columnas '*separación temporal*', 'cumple límite superior', '∄ *registro AVL intermedio*', '*cambio de grupo de embarque*' y '*grupo de embarque*', se han elegido para describir cómo se dividen los grupos de parada en grupos de embarque:

- Si una entrada es la primera de su grupo de parada ('*separación temporal*' = <nulo>), también debe comenzar un nuevo grupo de embarque (filas con '*rango sobre el conjunto*' ∈ {1,2,3,5, 6,8,9,45,46,47,49}). '*Cambio de grupo*' es igual a 1, y no es

necesario comprobar las columnas '*cumple límite superior*' o '*∄ registro AVL intermedio*'. Para cada una de estas filas, las columnas que intervienen en la identificación de su grupo de parada y grupo de embarque se rellenan con el mismo color, diferente de sus respectivas predecesoras.

- Si el lapso entre dos validaciones sucesivas del mismo '*grupo de parada'* es demasiado largo, son el final y el principio de dos '*grupos de embarque'* diferentes. La última fila muestra el símbolo ✗ en '*cumple límite superior*', mientras que su columna '*∄ registro AVL intermedio*' no es necesaria, y "*cambio de grupo*" es 1. También representa todo su proceso de decisión, utilizando un color para 'id parada', 'parámetro de agrupación' y 'grupo de paradas'; y otro para '*cumple límite superior*' e '*identificación del grupo de embarque*', mostrando cómo se divide cada '*grupo de paradas*' en grupos de embarque.
- Para el resto de pares de filas consecutivas que comparten el mismo '*grupo de parada'*, el símbolo de la columna '*∄ registro AVL intermedio*' indicará si pertenecen al mismo grupo de embarque:

  ✗: El vehículo relacionado con ambas entradas se ha movido a otra parada (y eventualmente ha vuelto) entre ambas lecturas de tiempo, por lo que pertenecen a diferentes grupos de embarque. De nuevo, '*cambio de grupo*'=1, y los colores ilustran el razonamiento de esta decisión: un color para '*grupo de parada*' y el primer '*grupo de embarque'*, y otro diferente para el segundo '*grupo de embarque*' creado por la división.

  ✓: No hay evidencia de que el vehículo se haya movido entre los instantes de ambas entradas, por lo que se concluye que pertenecen al mismo '*grupo de embarque*': '*cambio de grupo*' = 0. Las columnas de la última fila que deciden su '*grupo de parada*' y su '*grupo de embarque*' muestran los mismos colores que en la primera.

Se considerará que los grupos de embarque que duren más que el recorrido máximo para su ruta provienen de datos no fiables, y no se utilizarán para deducir las visitas perdidas a paradas no registradas por el IPTS.

2.2.1.3 Resultado

Los resultados del preprocesamiento del AFC se recogen en grupos_de_embarque, estructurada como se muestra en la tabla 3, mientras que la fig. 2 representa la transición de 15 eventos individuales de emisión de billetes a 4 grupos de embarque que engloban diferentes paradas.

Columna	*Tipo*	*Descripción*
id	entero	Id del grupo de accesos.
parada	entero	Id de la parada.
vehículo	entero	Vehicle id.
línea	entero	Route id.
grupo	entero	AFC group UID.
rang_embarq	rango temp.	[1ª validac., última validac.]

Tabla 3: Resultado del preprocesado de AFC: grupos_de_embarque

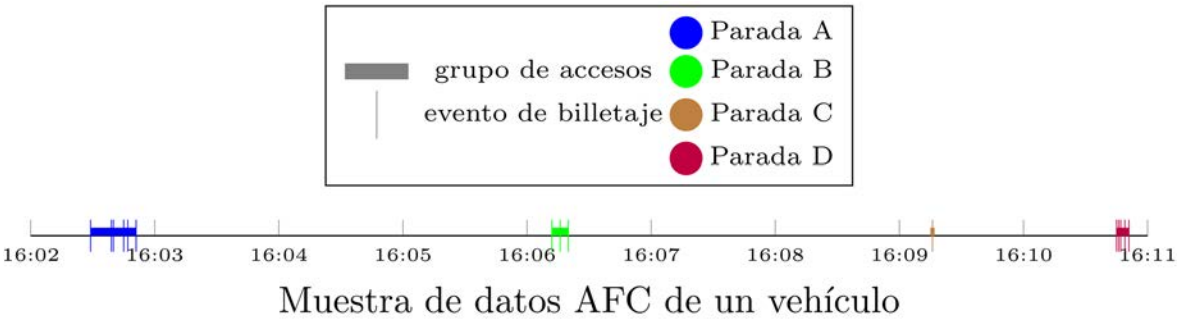


Figura 2: Recolección de eventos de billetaje individuales en grupos de embarque

2.2.2 AVL

Los procedimientos detallados en esta sección tienen como objetivo caracterizar el movimiento de los vehículos con un único registro para cada parada de cada servicio.

2.2.2.1 Eliminar filas duplicadas e inválidas

En primer lugar, se identifican y filtran las entradas duplicadas. Además, se supone que las entradas con un identificador de parada que no se corresponda con uno de los definidos en los datos de entrada (tabla 1-a), están causadas por eventos excepcionales que no corresponden a una llegada a una parada del sistema.

2.2.2.2 Identificar trayectorias

El siguiente paso es utilizar las columnas de los datos AVL para diferenciar las trayectorias que constituyen la oferta de transporte público. Para ello, en este trabajo se define una '*trayectoria*' como una serie de registros AVL consecutivos que comparten los mismos valores de *vehículo, línea* y *grupo*.(tabla 6a)

2.2.2.3 Determinar grupos de visitas

La tabla 4 muestra cómo se examina cada trayectoria para distinguir las ocasiones en las que se añade más de una fila al conjunto de datos para la misma escala en una parada (por ejemplo, cuando se vuelven a abrir las puertas para dejar entrar a un pasajero retrasado en el autobús). El procedimiento para identificar estos '*grupos de visitas'* (calcular los rangos de cada entrada sobre su trayectoria, y entre aquellos registros con los mismos valores de trayectoria y parada; y luego evaluar la variable de clasificación de cada elemento como su resta) es similar al que ya se ha descrito e implementado en 2.2.1.1 para encontrar *grupos de paradas* en los datos AFC.

Parada	*Rango cronológico en trajectoria*	-	*Rango cronológico en (traject., parada)*	=	*Número del grupo de visita*	*Parada*
A	1	-	1	=	0	A
B	2	-	1	=	1	B
C	3	-	1	=	2	C
C	4	-	2	=	2	
D	5	-	1	=	4	D
E	6	-	1	=	5	E
C	7	-	3	=	4	C
F	8	-	1	=	7	F
			...			

Tabla 4: Cómo se identifican los eventos de AVL vinculados a una única visita

2.2.2.4 Combinar los eventos de cada grupo de visitas

La información correspondiente a cada grupo de visita se resume en la tabla avl sintetizado (tabla 5), que incluye el instante más temprano y más tardío en el que según el AVL el bus estuvo en la parada.

instante	*duración*	*parada*
...		
12:50:29	0	151
12:51:11	19	135
12:51:18	198	135
12:51:47	<null>	135
12:55:07	14	134
...		

(a) *avl*

llegada	*salida*	*parada*
...		
12:50:29	12:50:29	151
12:51:11	12:54:36	135
12:55:07	12:55:21	134
...		

(b) *avl_sintetizado*

Tabla 5: Combinación de múltiples eventos AVL en una visita a la parada

2.2.2.5 Identificar y descartar desplazamientos entre paradas no plausibles

Considerando cada trayectoria como una serie de tramos de viaje entre sus grupos de visitas, no son posibles aquellos más cortos que el tiempo de viaje en flujo libre entre las paradas implicadas. Se dan dos situaciones posibles:

- Al retrasar la hora de salida en la parada anterior, aumentando así la longitud del tramo, se resuelve el problema. Esto equivale a suponer que la información sobre el tiempo que el autobús permaneció en la parada inicial del tramo no es fiable.
- Ni siquiera poniendo a cero el tiempo de permanencia en la primera parada hay tiempo suficiente para viajar a la segunda. En este caso, ambos grupos de visitas se considerarán poco fiables y se eliminarán.

Además, los tramos de viaje más largos que el límite superior establecido para su ruta (tabla 1f) se utilizarán para dividir sus trayectorias. Así, las entradas de AVL que presenten el mismo vehículo, ruta y grupo, pero separadas por un tramo de viaje demasiado largo para haber ocurrido durante un único servicio, se considerarán por separado.

2.2.2.6 Resultado

La tabla 6 muestra cómo se almacena el resultado del preprocesamiento AVL en las tablas *trayectorias* y *avl_sintetizado*.

Columna	*Tipo*	*Descripción*
id	entero	Id de la visita.
línea	entero	Id de línea.
vehículo	entero	Id vehículo.
group	entero	UID del grupo (AVL).
sec_paradas	entero	Id sec. paradas (descrita más adel.).
rang_trayect	rango de t.	[inicio tray., fin trayect.]

(a) *trayectorias*

Columna	*Tipo*	*Descripción*
id	entero	Id de la visita.
parada	entero	Id parada
ord_en_trj	entero	Orden cronológ. dentro de trayect.
trayectoria	entero	Id de trayectoria.
rango_avl	rango t.	[llegada, salida]

(b) *avl_sintetizado*

Tabla 6: Resultado del preprocesado de AVL

2.2.3 Planificación

En primer lugar, los eventos registrados en el subsistema de programación deben ser corregidos por el valor apropiado de ini_srv_crrct (tabla 1f, representado por *s* en la formulación de esta sección), si está definido para la ruta correspondiente.

A continuación, se crea un intervalo de tiempo *n* para cada servicio planificado, que engloba las horas de llegada y salida que pueden deducirse de las columnas más específicas disponibles, siempre que proporcionen una información coherente (por ejemplo: las salidas no pueden producirse antes que las llegadas). También se crea otro buffer de tiempo *q* en torno a su hora de inicio prevista t_p, con un semi-ancho igual al a la separación máxima entre servicios (tabla 1f, *s* en esta sección). Este intervalo se utilizará en la sección 3.7 para vincular cada entrada del horario con el servicio que lo materializa. Las ecuaciones (1) y (2) enuncian respectivamente el parámetro y las variables, y detallan las condiciones que se acaban de describir; mientras que la tabla 7 muestra la estructura de la información de planificación tras el preprocesamiento.

$$n = \left[[t_a, t_d]\ if\ t_a \leq t_d\right], \left[\left[t_p, t_d\right]\ if\ (t_a > t_d \vee \nexists t_a) \wedge t_p \leq t_d\right], \left[[t_d, t_d]\ if\ (t_a > t_d \vee \nexists t_a) \wedge t_p > t_d\right], [< null > en\ otro\ caso]] \quad (1)$$

$$q = [t_d - s, t_d + s] \quad (2)$$

donde:

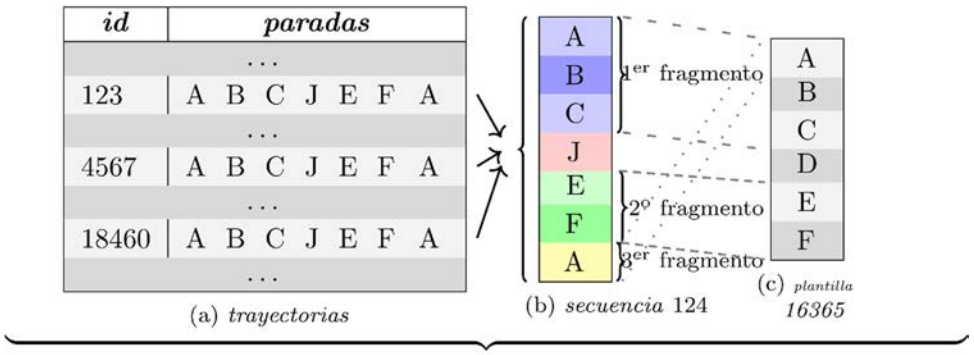
- tp: t. salida programado instante
- td: t. salida registrado instante
- ta: t. llegada bus registrado instante
- n: rango t. del subsist. De planif. [t. llegada, t. salida]
- q: búfer de búsqueda [cota t. inf., cota t. sup.]

Columna	*Tipo*	*Descripción*
id	entero	Id inicio serv.
parada	entero	Id parada.
vehicle	entero	Vehicle id.
buff_busq_serv	rango t.	[t. min, t. max] para búsqueda serv.
rango_planif	rango t.	[llegada, salida] de subs. planif.

Tabla 7: Resultado del preproceso de la planificación en bruto´

2.3 Analizar las trayectorias de la AVL como secuencias y fragmentos de rutas

Las trayectorias de AVL se analizan como secuencias ordenadas de paradas, que serán los bloques de construcción para ensamblar los servicios completos que se han producido, definidos por aquellas secuencias elegidas como ‘plantillas’.



Ordin. en sec.	*Ordin. en plantilla (núm. parada)*	*Longitud*	*Id sub-secuenc.*	*Id secuenc.*	*Plantilla*
1	1	3	1	124	16365
5	4	2	2	124	16365
7	1	1	3	124	16365

(d) *fragmentos*

Tabla 8: Análisis de las trayectorias de una línea, para una de sus plantillas

Columna	*Tipo*	*Descripción*
id	entero	Id de la secuencia.
línea	entero	Id de la ruta.
secuenc_parads	tupla de ent.	Secuenc. de ids de paradas

(a) *secuencias_paradas*

Columna	*Tipo*	*Descripción*
id	entero	Id de la secuencia.
n_stops	entero	Número de paradas.
name	texto	Nombre utilizado por los usuarios.

(b) *secuencias_plantillas*

Columna	*Tipo*	*Descripción*
ord_en_seq	entero	Ordinal del evento en su secuencia
núm_parada	entero	Ordinal del evneto en la plantilla.
fragmento	entero	Id del fragm. en su secuenc.
secuencia	entero	Id secuenc.
plantilla	entero	Id plantilla.

(c) *fragmentos*

Tabla 9: Resultado del análisis de las secuencias de AVL

La tabla 8 ilustra este proceso, y la 9 recoge los resultados de sus tres pasos.

2.3.1 Identificar distintas secuencias de trayectorias de la AVL

Se asigna un identificador a cada secuencia de paradas única extraída de las trayectorias de cada ruta, como se muestra en las tablas 8a, 8b y 9a. Esta relación se almacena en campo '*secuencia de paradas*' de la tabla de trayectorias.

2.3.2 Encontrar las secuencias a utilizar como plantillas

Esta metodología parte de la base de que cada línea puede dividirse en una serie de 'sublíneas' que representan los servicios que la componen (por ejemplo, los viajes de ida y vuelta entre los terminales; o un único viaje de ida y vuelta en el caso de las rutas circulares). Cada sublínea se caracteriza por su 'plantilla' de secuencia de paradas (tabla 8c) que debe seguir un recorrido típico, completo y perfectamente registrado de esa sublínea. Estas plantillas pueden conocerse de antemano, o averiguarse mediante el examen de las secuencias de paradas encontradas durante su paso anterior, y sus frecuencias relativas, ya que las plantillas estarán muy probablemente entre las que se encuentran más a menudo. Se almacenan como se ilustra en la tabla 9b.

2.3.3 Descomponer las secuencias en fragmentos de plantilla

Como se muestra en las tablas 8b a 8d, las secuencias seguidas por las trayectorias pueden dividirse en:

- Fragmentos continuos de las plantillas de su ruta (es decir, no faltan elementos entre sus extremos), que representan partes de servicios que el sistema AVL logró registrar correctamente. Permiten ver cada trayectoria encontrada en los datos AVL como una

serie de segmentos que se ajustan a su plantilla. El cuadro 9c muestra cómo se almacenan.

- Porciones incompatibles (causadas por entradas erróneas en el subsistema AVL; el vehículo que realiza otra ruta secundaria; u operaciones incorrectas, por ejemplo, no actualizar el ordenador de a bordo para reflejar que el autobús sigue una ruta diferente).

2.4 Elegir los modelos de distribución de los tiempos de recorrido y de parada

Estos modelos se utilizan como parte de los criterios para identificar los fragmentos de AVL o los grupos de embarque que forman parte del mismo servicio; para inferir la información de las paradas que faltan; y para filtrar los tiempos de inicio de servicio registrados erróneamente. Para cada ruta, se necesitarán los tiempos de viaje de los enlaces entre las paradas consecutivas y los tiempos de permanencia de todas ellas menos la última.

Deben tener en cuenta los factores conocidos que modifican los tiempos de desplazamiento y permanencia en la zona de estudio, como la hora, si es un día laborable o no, o los cambios de movilidad estacionales.

2.5 Ensamblar los servicios

Los servicios se construyen partiendo de una 'semilla' que se completa hacia atrás y hacia delante en el tiempo, buscando segmentos de AVL y eventos de grupos de embarque que formen parte de la misma sublínea y con el mismo identificador de vehículo que la semilla que, según el instante del punto de datos conocido más lejano en la dirección de crecimiento actual y las distribuciones de probabilidad de la duración de los tramos intermedios desconocidos del viaje y las llamadas en las paradas, caen dentro del intervalo de predicción de amplitud mínima de probabilidad ***g***.

'***g***' es un parámetro de esta metodología (ecuación 3). Cuanto más se acerque a uno, más exhaustiva será la búsqueda, y aumenta el riesgo de considerar eventos no válidos o no relacionados como parte de los servicios actuales. Sin embargo, si se fija demasiado bajo, pueden ignorarse eventos que realmente formaban parte del servicio que se está caracterizando.

$$g: probabilidad\ del\ intervalo\ de\ predicción \quad g \in [0,1] \tag{3}$$

Para cada sublínea y dirección (hacia atrás o hacia delante en el tiempo), las semillas se seleccionan siguiendo dos procesos iterativos consecutivos. En primer lugar, haciendo un bucle sobre los fragmentos de AVL con una longitud mínima ***c***, de mayor a menor. '***c***' es el parámetro "**longitud mínima de las semillas AVL**" (ecuación 4). Esta decisión parte de la hipótesis de que los fragmentos de trayectoria AVL más largos tienen más probabilidades de ser fiables, mientras que los más cortos pueden deberse a errores de reloj, GPS o de funcionamiento. A continuación, los grupos de embarque no filtrados se utilizarán también

como semillas. El algoritmo omitirá aquellas semillas contenidas en la tabla de eventos a ignorar (explicada en el último párrafo de esta sección).

$$c\text{: } Long.mín.de\ la\ semilla\ AVL \quad c \in \mathbb{N} - \{0\} \tag{4}$$

Una vez que se ha establecido una semilla, ésta "crece" tanto hacia atrás como hacia adelante, siguiendo un procedimiento que se asemeja al de la navegación por estima: partiendo del punto más lejano conocido en una dirección (el punto fijo inicial), se calculan intervalos de predicción de amplitud mínima de probabilidad $\boldsymbol{g}$ para las salidas o llegadas (si se recorre hacia atrás o hacia adelante, respectivamente) de las escalas en paradas consecutivamente más lejanas, como la suma de los tiempos de viaje y de parada de las paradas intermedias implicadas, hasta que se alcanza una de las siguientes condiciones (comprobadas en este orden) y se selecciona un nuevo punto fijo:

- El intervalo de predicción interseca el *rango_avl* de al menos un registro de la tabla *avl_sintetizado*. En este caso, se escoge el rango de horas de llegada y salida más cercano al más probable, y se identifica una porción del fragmento que lo engloba, para añadirlo al nuevo servicio en crecimiento, desde este registro hasta lo que se encuentre primero entre:

    - La penúltima o segunda parada de la ruta, mientras crece hacia adelante o hacia atrás, respectivamente.
    - El final de su fragmento en la dirección de crecimiento actual.

    Esta distinción tiene como objetivo, por un lado, ahorrar tiempo de computación, al sumar en un solo paso varias escalas del vehículo; y por otro, asegurar que siempre se calcule un rango de viabilidad en los terminales. Estos rangos, además de utilizarse como parte del proceso actual, para filtrar las entradas irreales del IPTS en esas paradas, se emplearán para decidir la mejor manera de incluir la información disponible de planificación.
- El intervalo de predicción se solapa con el rango de embarque de al menos un grupo de embarque compatible. Se elige el intervalo más cercano a las horas de llegada y salida más probables.
- Si la parada analizada es una terminal, se elige la llegada (o la salida, si se hace crecer el servicio hacia atrás) y el tiempo de permanencia más probables.

En la primera o en la segunda condición, "compatible" significa que se refiere a la misma ruta y al mismo vehículo que la semilla; y que no está en la tabla de eventos a ignorar (explicada en el último párrafo de esta sección). Si aparece más de una posibilidad, se selecciona la más probable según las distribuciones de tiempo de viaje y tiempo de parada.

En los tres casos, una vez seleccionado el nuevo punto fijo, el conjunto de valores más probables para los tiempos de viaje y los tiempos de permanencia del enlace se utilizará para inferir los tiempos de llegada y salida en las paradas intermedias que falten.

Después de llegar a una terminal, el crecimiento en la dirección actual termina. Para aquellas rutas en las que los datos en éstas se han considerado especialmente poco fiables (*term_dudos* = verdadero en la tabla 1f), si la llamada en la parada más cercana está respaldada por los datos de AVL o AFC, siempre se inferirán las horas de llegada y salida.

Una vez que una semilla ha crecido hasta abarcar un servicio completo, tal y como se describe en su plantilla; se crea un buffer que la engloba, extendiéndose hacia atrás y hacia delante en el tiempo desde la respectiva llegada y salida de cada llamada, añadiendo el límite inferior de tiempo de ida y vuelta para la ruta correspondiente (*t_vuelt_min*, tabla 1f). Los segmentos de AVL y los grupos de embarque con los que se superponga se añaden a las tablas de elementos a ignorar durante el resto del proceso de ensamblaje del servicio. Esto sirve para dos propósitos: hacer que ningún evento sea utilizado como parte de más de un servicio; y que los vehículos sigan itinerarios factibles (que pase suficiente tiempo antes de que vuelvan a la misma parada, como parte de otro servicio).

La figura 4 muestra un diagrama de flujo de la primera parte de este proceso, que utiliza segmentos de datos AVL como semillas. La segunda parte es completamente análoga, pero sólo se utiliza la información restante del AFC. La figura 3 ilustra un ejemplo completo, en el que los principales pasos son:

- (1): La semilla inicial es un segmento AVL que va desde la llegada a las :20:13 a AB, hasta la salida a las :22:41 de AE.
- (2): Ésta crece hacia atrás, utilizando el intervalo de búsqueda [:18:51, :19:53] en el punto final AA. Se ha definido fijando la llegada a AB como punto fijo, y calculando el intervalo de predicción con probabilidad $\boldsymbol{g}$ para la presencia del vehículo en AA.
- (3): Se encuentra un único evento AVL compatible y superpuesto (3a), con horas de llegada y salida :18:31 y :18:55, respectivamente:

    - Si las lecturas en los terminales de esta ruta se han considerado tan fiables como en otras paradas (y = falso), se aceptará (3a) como la llamada del autobús en el terminal inicial.
    - En caso contrario, dado que el punto fijo del que parte la búsqueda se encuentra en la parada próxima al terminal (3c), se preferirá la visita inferida 3b, de :19:15 a :19:46.

  Al ser uno de los extremos de la ruta, el crecimiento hacia atrás termina.

- (4): A medida que se avanza, se calcula el intervalo de búsqueda que se utilizará en AF, utilizando como referencia fija la hora de salida de AE (:22:41). El resultado es el intervalo de predicción de probabilidad g de la presencia del autobús en AF: [:23:04, :24:05], que no cruza ninguna entrada compatible de los subsistemas AVL o AFC.
- (5): Se sigue buscando hacia adelante. En AG, se crea otro intervalo de predicción con probabilidad ***g*** para la llegada del autobús. Esta vez, se necesitará la suma de las distribuciones individuales de los tiempos de viaje de AE a AF, y de AF a AG; y del tiempo de parada en AF. El rango resultante ([:23:17, :24:57]) se solapa con un grupo de embarque ([:23:55, :23:59]). Sus eventos de embarque más tempranos y más tardíos se utilizarán como aproximación a la llegada y salida en AG.
- (6): Considerando ahora el intervalo de 1m14s entre la salida del autobús de AE a las :22:41, y la llegada a AG a las :23:55; la combinación más probable de los tiempos de viaje de AE a AF y de AF a AG; y del tiempo de parada en AF es, según sus respectivas distribuciones probabilísticas, 45s, 26s y 3s; respectivamente. Así, los tiempos de llegada y salida en AF se fijan en [:23:26, :23:29].
- (7): De nuevo, la búsqueda se realiza en la parada H. Esta vez, se encuentra una entrada AVL compatible. Ésta y otras tres del mismo fragmento se añaden al servicio.
- (8): Hubo que deducir varias paradas intermedias entre la salida de AK y la llegada y AR. Las horas de llegada y salida que faltan se fijarán en sus valores más probables, de acuerdo con las 7 distribuciones de tiempo de viaje y 6 de tiempo de permanencia implicadas.
- (9): Por último, se llega al otro extremo de la ruta. Dado que no se encuentra ningún AVL o AFC compatible, la llegada a esta parada; así como las llegadas y salidas a otras posteriores a la última salida conocida, si la hay; se fijan en sus valores medios.

El resultado después de utilizar todos los datos de AVL y AFC (tabla 10) es el resultado de esta metodología, y consiste en tres tablas:

- *servicios*, que sintetiza cada uno de los servicios que han sido detectados por esta metodología.
- *visitas_a_paradas*, que caracteriza a cada servicio.
- *rangos_búsqueda*, donde se guardan los intervalos de predicción utilizados durante la creación de los servicios.

Columna	*Tipo*	*Descripción*
vehículo	entero	Id vehículo.
id	entero	Id servicio.
serv_comb	entero	Id del serv. que lo engloba (si es aplic.).
plantilla	entero	Id. plantilla
comienz_prog	entero	Id del inicio planeado (si es aplic.).
rang_serv	rango t.	[Salida 1ª parada, llegada a últ. parada]
servs_comb	tupla de ent.	Ids de servicios englobados (si es aplic.)

(a) *servicios*

Columna	*Tipo*	*Descripción*
núm_parada	entero	Ordinal of stop in template.
id	entero	Id dc la visita.
servicio	entero	Id del servic. del que forma parte.
avl_sint_id	entero	Fuente en *avl_sintetizado* (si es aplic.).
gr_acc	entero	Fuente en *grupos_de_accesos* (si es aplic.).
rango_visita	rango t.	[llegada, salida]

(b) *visitas_a_paradas*

Columna	*Tipo*	*Descripción*
origen	entero	Ordinal de la última parada con información del IPTS.
num_parada	entero	Ordinal de la parada analizada.
servicio	entero	Servicio del que forma parte la visita.
rang_búsq	rango t.	Extremos del intervalo de predicción.

(c) *rangos_búsqueda*

Tabla 10: Resultado de la caracterización de servicios

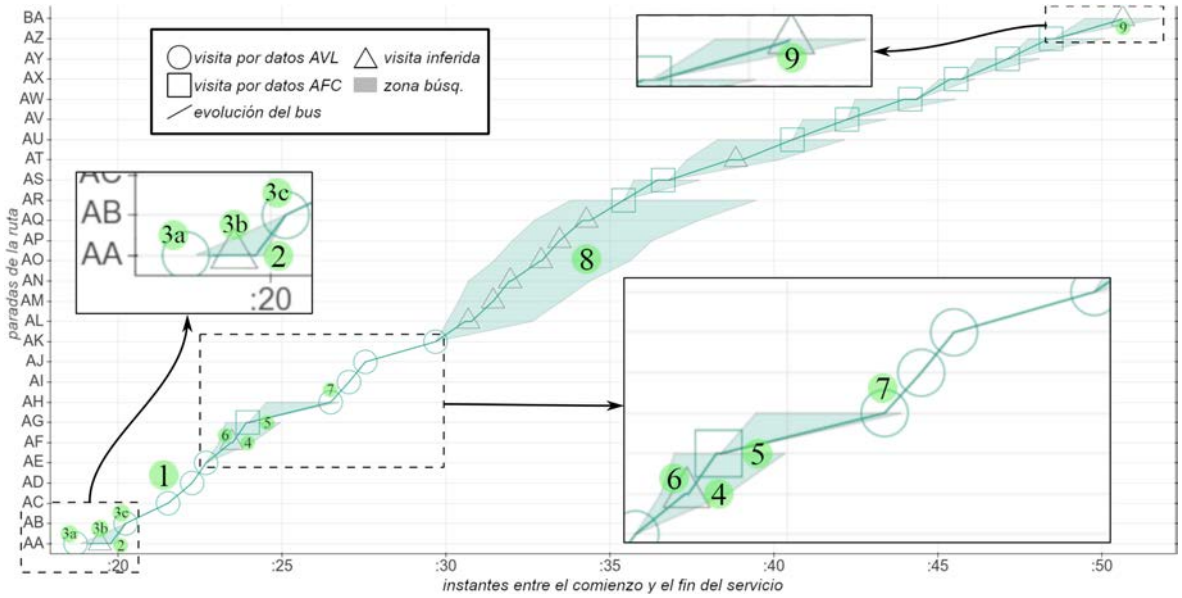


Figura 3: Proceso de inferencia de un servicio

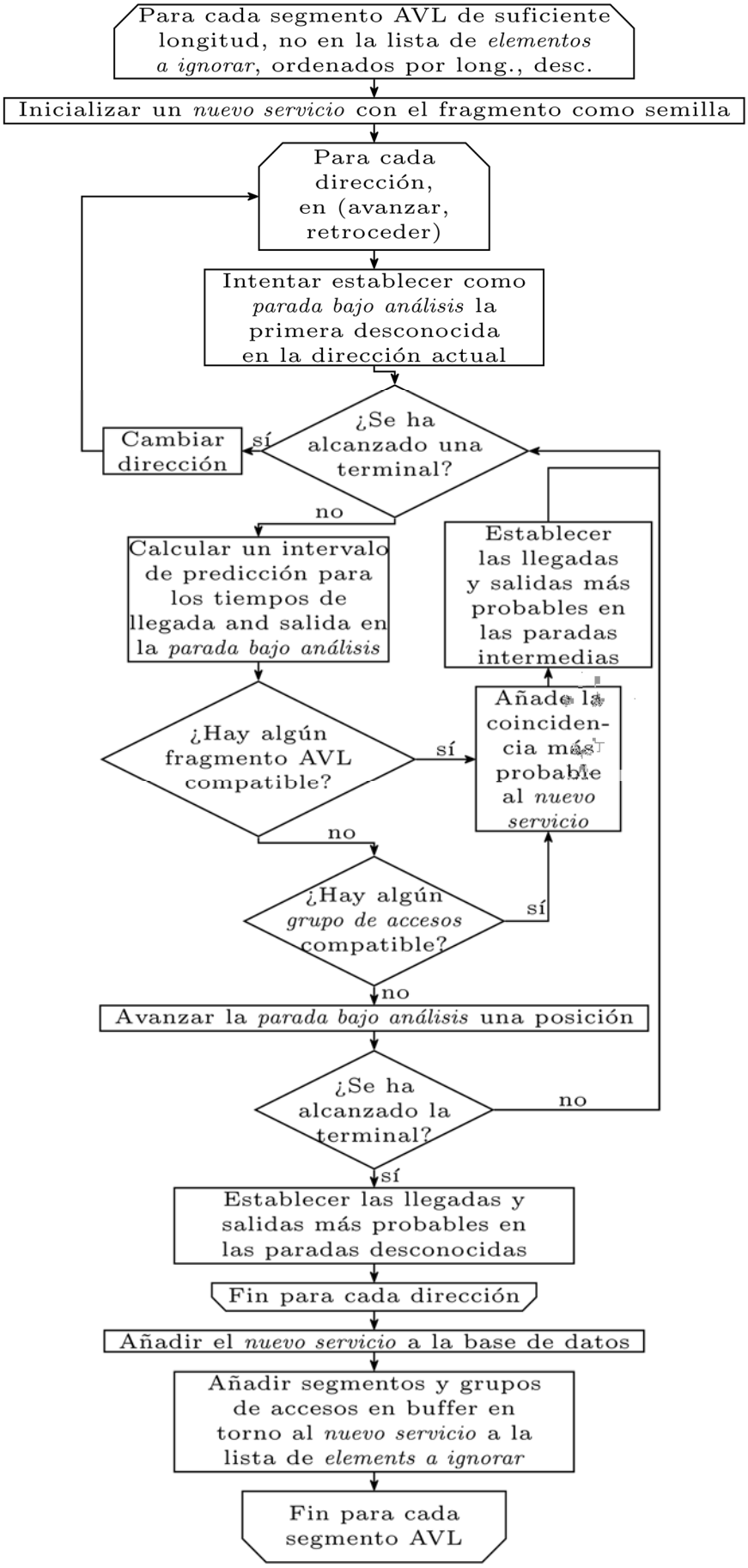


Figura 4: Proceso de inferencia de un servicio a partir de una semilla AVL

2.6 (Opcional) detectar y fusionar los casos en los que un vehículo ha cambiado su id a mitad de servicio

Algunos de los servicios del caso presentado en la siguiente sección de este trabajo varían su identificación de vehículo a mitad de servicio una vez. Pueden detectarse en esta metodología como dos servicios, uno 'anterior' y otro 'posterior', extremadamente cercanos en el tiempo; en los que un único vehículo podría haber proporcionado todas las visitas no inferidas a las paradas. Esta sección sigue la nomenclatura descrita en (5).

$[\emptyset, \lambda$:ids de servs. ant. y post. con $\emptyset, \lambda \in \mathbb{Z}]$, $[C$:relación deben combinarse con $C = \{(\emptyset, \lambda)/\emptyset, \lambda$ son el mismo serv.$\}]$, $[T$:instantes. son los posibles instantes$]$, $[R$:rangos t. con $R = \{(\rho[0], \rho[1]) \in T^2, \rho[0] \leq \rho[1]\}]$, $\left[\&\&$:relación "se superponen" con $\&\& = \{\rho_i, \rho_j\}, \rho_j[0] \leq \rho_i[1] \leq \rho_j[1] \vee \rho_j[0] \leq \rho_i[0] \leq \rho_j[1] \vee \rho_i[0] < \rho_j[0] \wedge \rho_i[1] > \rho_j[1]\right]$, $[\sigma_i$:rango del serv. i-ésimo con $\sigma =$ (comienzo, fin) $\in R]$, $\left[\upsilon_{i,j}$: visit range for service i at stop j con $\upsilon =$ (llegada, salida) $\in R\right]$, $\left[\in_{i,j}$: rango de búsqueda del servicio i en parada j con $\in =$ (cota inf., cota sup.) $\in R\right]$, $[\tau$: núm. parada con $\tau \in \mathbb{N}]$, $[\sigma\mu_i$:cota inf. del t. viaje entre paradas i e $i+1]$ (5)

Para corregir este problema, los autores proponen el siguiente procedimiento que se debe llevar a cabo para cada secuencia utilizada como plantilla de servicios

2.6.1 Identificar los pares de servicios que deben combinarse

- Para ahorrar tiempo de cálculo, sólo se considerarán aquellos servicios que presenten solapamiento entre sus correspondientes rangos de servicio:

$$\emptyset C \lambda \Longrightarrow \sigma_\emptyset \&\& \sigma_\lambda \quad (6)$$

- Además, para cada escala de cada servicio en una parada, se crea un buffer de tiempo, como el más pequeño que incluye su rango de visita y, si existe, su rango de búsqueda. Se considera que dos servicios se suceden lo suficientemente cerca como para ser candidatos viables cuando sus buffers de tiempo se solapan.

$$\emptyset C \lambda \Longrightarrow \exists \tau, \upsilon_{\emptyset,\tau} \&\& \upsilon_{\lambda,\tau} \vee \in_{\emptyset,\tau} \&\& \in_{\lambda,\tau} \vee \upsilon_{\emptyset,\tau} \&\& \in_{\lambda,\tau} \vee \in_{\emptyset,\tau} \&\& \in_{\lambda,\tau} \quad (7)$$

- Por último, debe ser posible, teniendo en cuenta los límites inferiores de los tiempos de viaje entre paradas, que un solo autobús realice todas las entradas de visitas a paradas de ambos servicios que se desprenden de los datos del IPTS. El cumplimiento de esta condición depende del número de parada más alto para el que el servicio 'anterior' presenta una entrada de visitas a paradas no inferida (τ_φ); y, correspondientemente, del más bajo del 'posterior' (τ_λ):

- Si $\tau_\varphi = \tau_\lambda = \tau$, ambos representan la misma parada, en la que el IPTS tiene registros con los identificadores de los vehículos antiguos y nuevos. En dicha parada se calculan los siguientes rangos de tiempo:

  - Un "intervalo de viabilidad" $\zeta_{\varphi,\lambda}$ que delimita el lapso de tiempo en el que es posible que el autobús haya llegado tras salir de la $(\tau\text{-}1)^{\text{ésima}}$ parada, como se describe en el servicio 'anterior' φ, y aun así llegue a la $(\tau+1)^{\text{ésima}}$ del 'posterior' λ, teniendo en cuenta los límites mínimos de las duraciones de los tramos de viaje implicados:

$$\xi_{\emptyset,\lambda} \in R;\ \xi_{\emptyset,\lambda} = (v_{\emptyset,(\tau-1)}[1] + \mu_{(\tau-1)}, v_{\lambda,(\tau+1)}[0] - \mu_\tau) \quad (8)$$

  - Un "rango de presencia de bus" $\eta_{\varphi,\lambda}$, que es el rango de mínimo-ancho que engloba a los de los servicios 'anterior' y 'posterior':

$$\eta_{\emptyset,\lambda} \in R;\ \eta_{\emptyset,\lambda} = (min(v_{\emptyset,\tau}[0], v_{\lambda,\tau}[0]), max(v_{\emptyset,\tau}[01], v_{\lambda,\tau}[1])) \quad (9)$$

    La condición se cumple si estos dos rangos se superponen:

$$\emptyset C\lambda \wedge \tau_\emptyset = \tau_\lambda \Longrightarrow \xi_{\emptyset,\lambda} \&\& \eta_{\emptyset,\lambda} \quad (10)$$

- Si $\tau_\varphi < \tau_\lambda$, el lapso entre la salida registrada del primer servicio desde la parada τ_φ y la llegada registrada del segundo a τ_λ debe ser mayor o igual que el límite inferior del tiempo total de viaje entre ellos.
- Si $\tau_\varphi > \tau_\lambda$, los dos servicios candidatos no proceden de uno solo que haya cambiado su id.

La ecuación (11) resume estos criterios:

$$\emptyset C\lambda \Leftrightarrow \{[\xi_{\emptyset,\lambda} \&\& \eta_{\emptyset,\lambda}\ if\ \tau_\emptyset = \tau_\lambda], [v_{\lambda,\tau_\lambda}[0] - v_{\emptyset,\tau_\emptyset}[1] \leq \sum_{\tau=\tau_\lambda}^{\tau_{\emptyset-1}} \mu_i\ \ if\ \tau_\emptyset < \tau_\lambda\} \quad (11)$$

2.6.2 Actualizar las tablas de caracterización de los servicios

Los servicios que cumplen las ecuaciones. (6), (7) y (11) se fusionan en uno nuevo. Las horas de llegada y salida de cualquier parada entre τ_φ y τ_λ se elegirán como las más probables, según las distribuciones de tiempo de permanencia y de viaje; y la información para relacionarlos se almacena en las columnas serv_comb y servs_comb. de la tabla 10a.

La figura 5 ilustra un ejemplo, en el que la metodología detecta que las entradas que en una primera aproximación se utilizaron para afirmar que dos servicios diferentes de una ruta entre las paradas AR y BJ tuvieron lugar (azul y naranja) son en realidad parte de uno solo,

y entonces reevaluar las llamadas desconocidas en las que este hecho puede ser utilizado para mejorar las estimaciones de llegada y salida:

- El estado de los dos servicios propuestos por el apartado 2.5 de esta metodología cumple con las condiciones que los identifican como uno solo, con un cambio de id de vehículo intermedio:

    - Sus buffers temporales se solapan en al menos una parada, como puede verse observando las partes coloreadas en azul y naranja.
    - Considerando sólo las escalas respaldadas por datos del IPTS, la más tardía de uno de los servicios (visita del servicio anterior en AY, 17:11:50, marcado 1a) ocurre en una parada anterior a la más temprana del posterior (visita a BE, 17:22:05, marcada 1b). El intervalo entre la salida del primero y la llegada al segundo es de 10m15s, mientras que la suma de los límites inferiores de los tramos de viaje implicados es de 2m23s, lo que significa que un solo vehículo podría ser responsable de ambos.

- Se recalculan los tiempos intermedios de llegada y salida entre AY y BE. En lugar de sus valores medios según sus respectivas distribuciones y la salida de AY o la llegada a BE; adoptarán la combinación más probable de valores que satisfagan ambas condiciones al mismo tiempo.

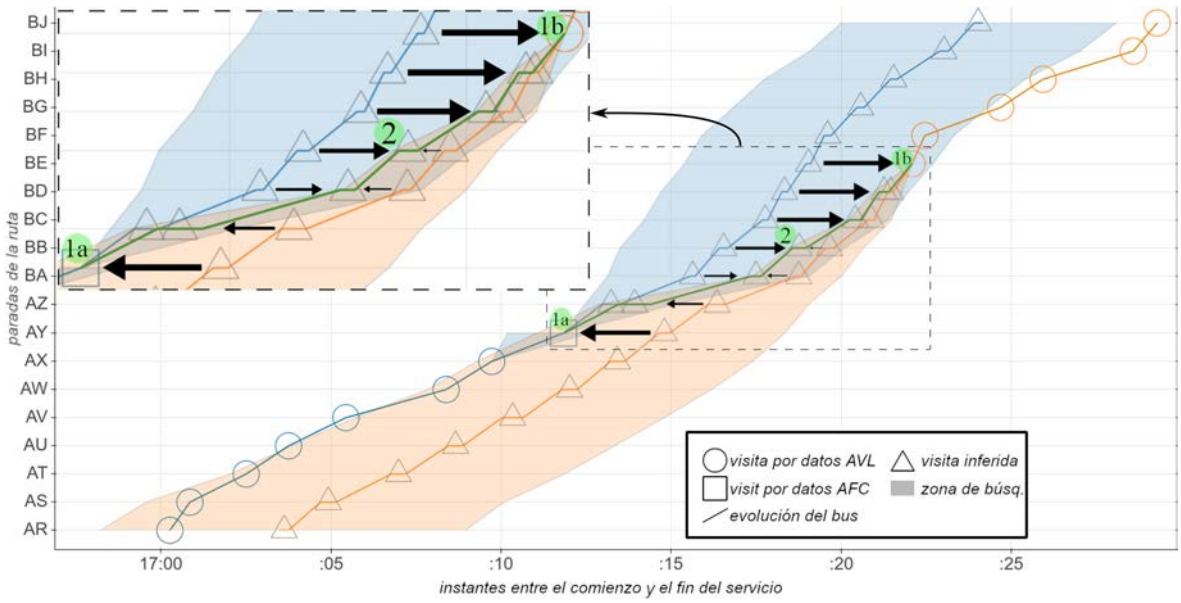


Figura 5: Fusión de 2 servicios (‘anterior’ azul, ‘posterior’ naranja, visitas modificadas: línea verde)

2.7 Asignación de servicios a los recorridos programados y actualización de los intervalos de tiempo de las visitas

Esta parte de la metodología tiene varios objetivos: en primer lugar, diferenciar entre los servicios previstos que se materializaron o no; identificar los recorridos extra no programados; y eliminar las visitas inferidas a paradas que no se produjeron realmente, para aquellos servicios que se identifican con éxito como que comienzan ‘aguas abajo’ de la terminal inicial.

Una vez que se ha vinculado un servicio con su inicio programado, la información adicional de la tabla de horarios puede utilizarse para afinar aún más las horas de llegada y salida. Estos son los pasos propuestos, que también se muestran en forma de diagrama de flujo en la fig. 6:

- Se realiza un bucle sobre todos los pares (inicio programado, servicio) en los que la salida de este último de la parada prevista cae dentro del buffer *q* (eq. 2) del primero, dando prioridad a los pares que que comparten la mismo *id de vehículo*, y luego ordenados por el valor absoluto del lapso entre la salida del servicio y el inicio programado, de forma ascendente. A menos que alguno de ellos ya haya sido vinculado, pasan a serlo entre sí.
- Si se encuentra un emparejamiento, a partir del terminal inicial de toda la ruta, se eliminan consecutivamente las visitas inferidas a las paradas, hasta llegar a una que esté respaldada por registros AFC o AVL.
- Si el subsistema de planificación registró el inicio del servicio, se evaluará la verosimilitud de su correspondiente rango de tiempo *n* (eq. 2), utilizando el rango de viabilidad apropiado almacenado en la tabla de rangos de búsqueda (tabla 10c, si no está disponible, se crea uno utilizando la llamada más cercana del servicio con soporte de datos). Si *n* se considera creíble, pueden darse dos situaciones:
    - Si la escala del servicio se había deducido previamente (sección 2.5) a partir de otros datos del IPTS, la información disponible se combinará para obtener las presencias más temprana y más tardía del autobús en esa parada.
    - En caso contrario, se utilizará n como rango [llegada, salida] al inicio del servicio.
- Las visitas inferidas aguas abajo, hasta la primera sustentada por datos del IPTS, se mejoran hasta sus nuevos valores más probables, considerando el tiempo total de viaje entre el inicio del servicio programado y ese primer punto de datos conocido, y las distribuciones de tiempo de viaje y de parada.

La figura 7 muestra las primeras paradas de un servicio de ejemplo:

- Su estimación inicial se ha vinculado a una salida prevista en la parada AF, con una diferencia entre sus tiempos de salida inferidos y los previstos de 51s.
- Las paradas anteriores al inicio previsto no están respaldadas por ningún registro IPTS, y se borran.
- En este caso, la llegada y la salida fueron registradas por el subsistema de planificación a las 07:25:39 y 07:26:01, respectivamente. Como estas horas están dentro del rango de búsqueda de ese servicio en la parada AF ([07:23:39, 07:26:13]), se aceptan como lo que realmente ocurrió.
- Las visitas a AG, AH, AI y AJ también se recalculan teniendo en cuenta la nueva información.

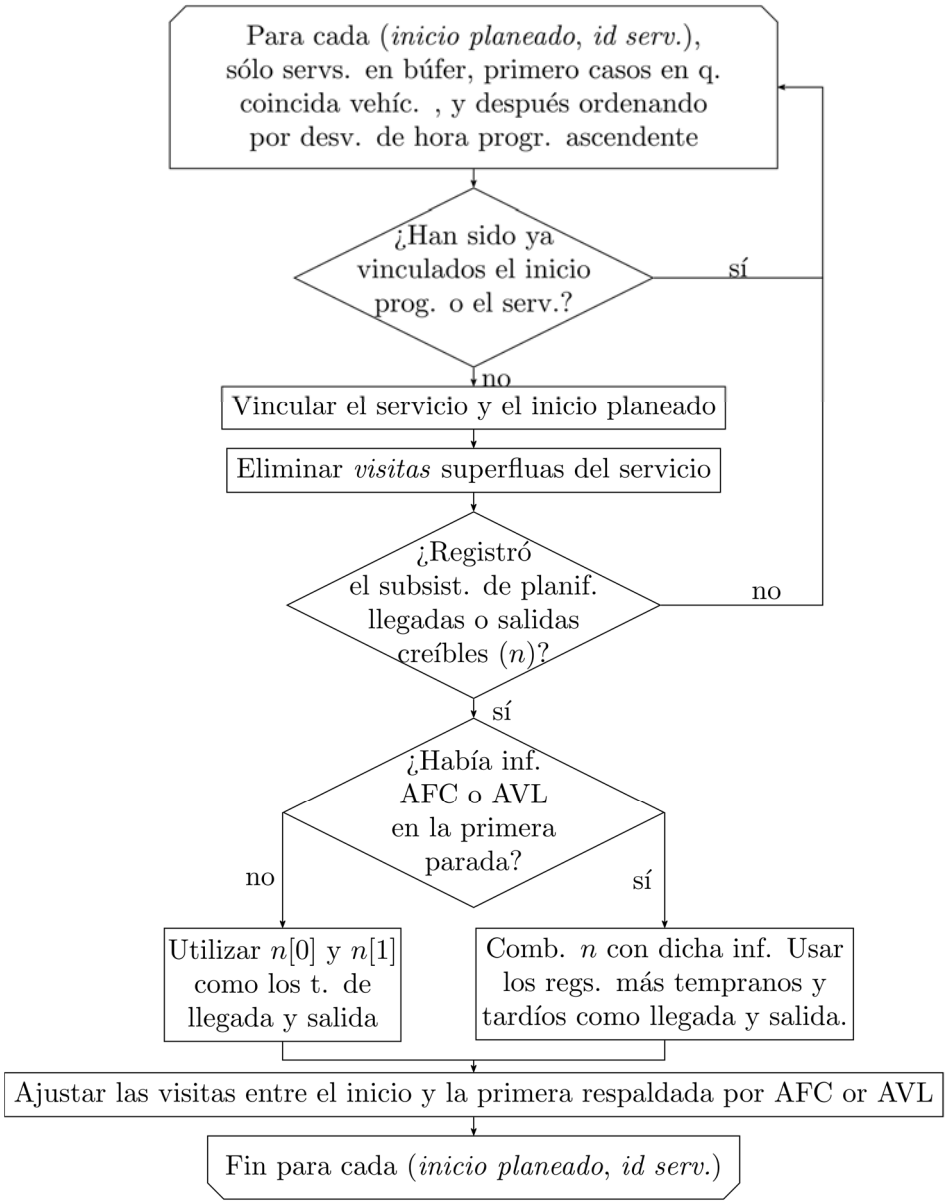


Figura 6: Asociación de servicios programados e inferidos

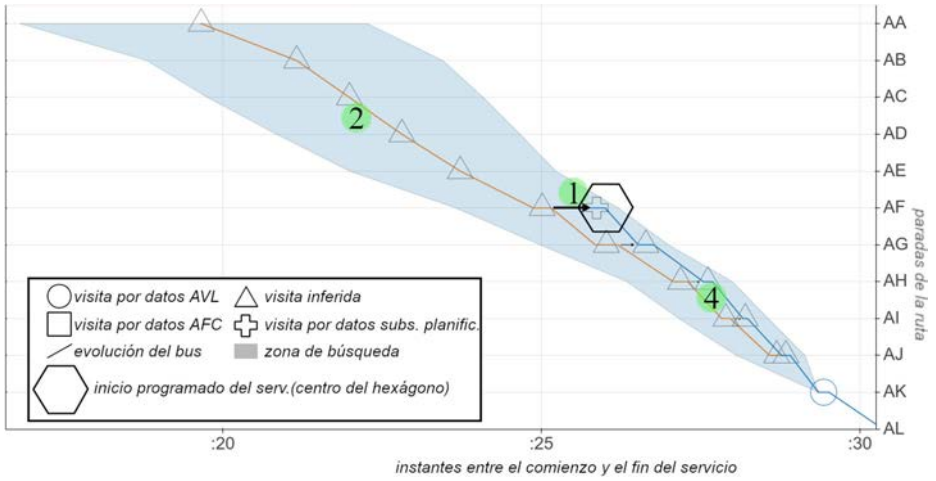


Figura 7: Mejora de la llegada y salida de un servicio una vez identificado su probable inicio programado. Caracterización final en azul, entradas modificadas en naranja

2.8 Imputación de grupos de embarque

Una vez definidas y refinadas las escalas de todos los servicios posibles, los grupos de embarque se asignarán primero a un servicio y luego a la parada en la que se produjeron.

Para la primera tarea, un rango de imputación (ligeramente resaltado con un patrón diagonal para el segundo caso de la fig. 8) se crea para cada servicio desde el momento en que el vehículo llegó a su parada inicial, menos el límite superior del tiempo entre servicios (‘*s*’ *en la fig. 7*, *sep_max* en la tabla 1f) para asegurarse de que se identifican todos los eventos AFC pertinentes en el terminal inicial, marcado como 6a); hasta el momento en que dejó su penúltima parada (ya que no se deben asignar eventos AFC a la última parada de un servicio), más el parámetro ‘***o***’ (), que permite cierto margen entre los eventos AVL y AFC, para cubrir casos como validaciones después de que el vehículo abandone la parada o desincronizaciones menores del reloj, marcado como 6b.

$$\boldsymbol{o}: \text{holgura AFC} \quad \text{tiempo} \quad (12)$$

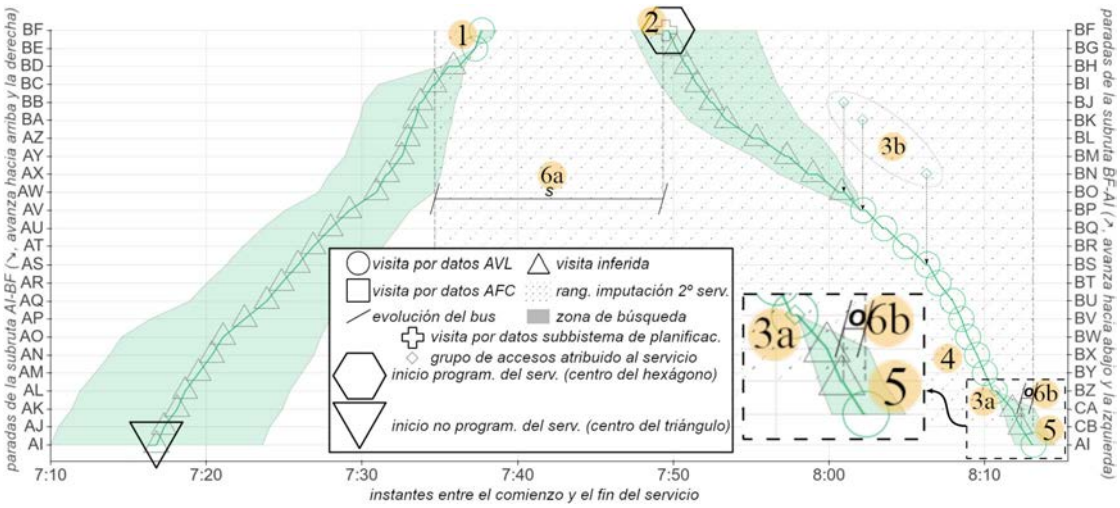


Figura 8: **Imputación de los grupos de embarque y análisis de los datos del IPTS**

Cada *grupo de embarque* se trata en un proceso de dos pasos:

- En primer lugar, se asigna al servicio cuya franja horaria solapa, y que se refiere al mismo vehículo y ruta. Si no se encuentra ningún servicio, se ignora el requisito de coincidencia de ruta, para tratar aquellos casos en los que el estado del subsistema de expedición de billetes no reflejaba la ruta que seguía realmente el vehículo. Cualquier grupo de embarque que quede no estará vinculado a un servicio.
- A continuación, se identifica la parada adecuada dentro del servicio, considerando todas sus escalas menos la última:

    - Si la diferencia entre el *rango_embarque* (tabla 3) y el *rango de visita* (tabla 10a) en la parada especificada por el grupo de embarque es menor o igual a la holgura ***o***, esa parada será aceptada como aquella en la que los viajeros subieron al autobús (por ejemplo, 3a en la fig. 8).
    - En caso contrario, se asumirá que el AFC no identificó correctamente la identificación de la parada. En su lugar, se elegirá el de la llamada más cercana del vehículo (por ejemplo, 3b en la fig. 8, donde los 3 grupos de embarque que se registraron como ocurridos en las paradas BJ, BK y BN se asignan respectivamente a BO, BP y BS).

2.9 Seleccionar servicios respaldados por suficiente información

El último paso es establecer y aplicar un criterio para aceptar o rechazar cada uno de los posibles servicios que han sido identificados por esta metodología. Se sugiere establecer límites que consideren estas características (ecuación 13):

- Si se asignó o no una salida planificada al servicio (*w*). En este último caso, también hay que tener en cuenta si la identificación del vehículo es la misma en ambas bases de datos (*p*), y si el subsistema de programación registró una hora de salida compatible (*v*).
- El número total *h* de grupos de embarque atribuidos al servicio, como se ha descrito en la sección 2.8.
- Cuántas visitas de ese servicio provienen de la información de AVL (*f*).
- El número de paradas entre la primera y la última visita apoyada por los datos del IPTS (*l*).

[w: el servicio esta planeado (lógico)], [p: se empleó el vehículo previsto (lógico)],
[v: Subs. de planif. resgistró un t. salida válido (lógico)],
[h: Núm. de grupos de acceso ($h \in \mathbb{N}$)], [f: Visitas respaldadas por AVL ($f \in \mathbb{N}$)],
[l: separación máx. entre visitas respaldadas por IPTS ($l \in \mathbb{N}$)] (13)

La figura 8 ofrece un ejemplo, analizando dos posibles servicios consecutivos de un vehículo, que cubren sub-líneas complementarias entre los terminales AI y BF, ambos

compuestos por 23 tramos de viaje. Sólo 2 entradas consecutivas de la tabla *avl_sintetizado* apuntan a la existencia del primero (1); mientras que el segundo se apoya en un servicio planificado de ese vehículo para el que el subsistema de programación registró la primera escala (2), 4 grupos de embarque (mostrados en 3a y 3b), 12 filas de *avl_sintetizado* (4), y por el hecho de que el intervalo entre la primera (2) y la última (5) escala, obtenidas a partir de observaciones registradas en el IPTS cubre toda la ruta. Es casi seguro que el primer servicio no ocurrió, mientras que el segundo es muy probable que sí.

3. APLICACIÓN A UN CASO REAL

Los resultados de esta metodología se ilustran utilizando los eventos AVL y AFC, y los inicios de servicio programados de los vehículos que, durante 1 año, recorren la línea 1 en Santander, una ciudad de la costa norte de España (fig. 9).

Figura 9: Ciudad de Santander [mapa, Wikimedia] [foto, SPIEGEL]

Funciona aproximadamente desde las 07:00 hasta las 23:00, con intervalos entre servicios de como máximo s = 20min. En aproximadamente la mitad de las ocasiones, el subsistema de programación registra, con una desviación de alrededor de z = 20s, la llegada y salida del vehículo de la primera parada del servicio. Un viaje completo de ida y vuelta requiere al menos d = 1h, mientras que un tramo de viaje único, incluso en las circunstancias más desfavorables, no debería durar más de e = 15min. Aunque el IPTS es extremadamente útil durante las operaciones diarias, la explotación de sus datos tiene que superar varios problemas:

- Baja fiabilidad de AVL y AFC en la mayoría de los extremos de los servicios (y = verdadero), debido a la forma en que a veces se manejan los ordenadores de abordo y al hecho de que cuando un autobús está vacío al acercarse al final de la ruta, los conductores a menudo encuentran más conveniente esperar hasta su próximo recorrido en una parada anterior a la final.
- Diariamente, cada viaje que cubre una de las 2 sub-líneas a veces no puede ser identificado de forma fiable con un *id* dentro de los conjuntos de datos AVL y AFC: este campo puede mostrar varios valores dentro de un mismo viaje, o el mismo valor puede ser utilizado para recorridos consecutivos que cubren ambas sub-líneas.

Además, este identificador no es coherente entre las informaciones de AVL, AFC y planificación.

- Falta de entradas AVL.
- Eventos AVL y AFC erróneos que se derivan de las limitaciones del IPTS, como la pérdida de la señal GPS, los fallos de comunicación o los errores del ordenador de a bordo; o de operaciones atípicas o incorrectas (por ejemplo, el establecimiento de parámetros de estado del vehículo que identifican erróneamente la tarea que se está realizando).
- La información relativa a si un viaje planificado finalmente ocurrió y cuándo comenzó es la mayoría de las veces precisa, pero a veces el inicio de un servicio normalmente realizado no se registra, o lo hace con marcas de tiempo muy inexactas.
- En ocasiones, al cambiar el conductor de un bus en mitad de un servicio, el identificador del vehículo cambia, por lo que presentarán 2 valores diferentes.

Esta implementación utiliza el lenguaje procedimental PL/pgSQL dentro de una base de datos PostgresSQL 13.2 para sus tareas principales; y Python 3.8 y Bokeh 2.2 para mostrar una representación interactiva de los resultados.

3.1 Aplicación de la metodología

3.1.1 Datos de entrada

3.1.1.1 Paradas y sub-líneas

Santander cuenta con aproximadamente 460 paradas de autobús. La ubicación de las 75 que conforman la ruta 1, que se divide en dos sub-lineas con una parada intermedia ("Consuelo Berges 16") y ambas terminales en común, se muestra en la fig. 10. Estas sub-líneas proporcionan las plantillas que se utilizarán para desglosar las secuencias de paradas encontradas durante el tratamiento de los datos de la AVL.

Figura 10: Paradas de la línea 1

Este itinerario comienza en el parque científico PCTCAN, en el oeste, y atraviesa la ciudad hacia el este a través de las principales arterias, pasando por muchos de sus centros comerciales, residenciales, turísticos y administrativos hasta llegar al Parque de la Península de La Magdalena (uno de sus principales lugares de ocio). A continuación, gira hacia el noroeste y sigue el litoral, dando acceso a las playas más populares de Santander. Finalmente desemboca en Valdenoja, un barrio que aunque presenta algún uso comercial limitado se puede caracterizar como un suburbio dormitorio.

Durante los días no laborables la actividad en el PCTCAN disminuye considerablemente, por lo que los autobuses no visitan las 3 paradas más orientales. Además, especialmente durante los días laborables, varios servicios de refuerzo planificados pero no anunciados comienzan “aguas abajo” por la primera parada, para aprovechar espacios libres breves que tengan los conductores entre otras asignaciones.

3.1.1.2 AFC

El conjunto de datos incluye 2586600 eventos AFC en bruto. Casi todos (99,99%) corresponden a paradas reales dentro de la ciudad, mientras que el resto tiene ids que no se refieren a una parada física.

3.1.1.3 AVL

Hay 1569417 eventos AVL en bruto. Todos representan llamadas en paradas reales de la ciudad.

3.1.1.4 Información del subsistema de planificación

Mientras que el horario diario que los viajeros tienen en cuenta a la hora de planificar sus viajes en la ruta 1 especifica, dependiendo de si es un día laborable o no, alrededor de 100 u 80 lugares y horas en los que comienza un servicio, la autoridad de transporte planifica algunos recorridos extra de vehículos reales, ofreciendo servicios adicionales menos conocidos de la ruta, como varios que comienzan en el hospital Valdecilla para el personal que acaba de terminar sus turnos, o reforzando la oferta durante los períodos conocidos de máxima demanda cuando la distribución de los recursos disponibles lo permite. Los recorridos adicionales de los vehículos que no están presentes en la información de programación pueden producirse debido a decisiones tácticas durante las operaciones diarias.

3.1.2 Preprocesamiento

3.1.2.1 AFC

Siguiendo la metodología expuesta en el apartado 2.2.1, se han encontrado 719971 grupos de paradas., divididos en 724550 grupos de embarque (un 0,6% más de eventos). De ellos, 108 (0,01%) duran más de *s* y no se tendrán en cuenta.

Hay, en promedio, 1 grupo de embarque por cada 4 eventos de billetaje en bruto. Estos grupos de embarque proporcionan además una primera estimación de reserva de las horas de llegada y salida en las paradas, que se utilizará si no se dispone otros registros.

3.1.2.2 AVL

Como se explica en el apartado 2.2.2, los eventos AVL consecutivos que representan la misma visita a una parada se fusionan, dejando 1532299 entradas (un 2% menos). De ellas, 78520 (5%) se consideran poco fiables porque forman parte de tramos de viaje imposiblemente cortos. Las 1453779 entradas restantes, reunidas en la tabla *avl_sintetizado*, se clasifican en 45840 trayectorias.

3.1.3 Analizar las trayectorias de la AVL como secuencias

Las 45840 trayectorias presentan 5800 secuencias de paradas diferentes. Las dos más frecuentes coinciden con los itinerarios ya conocidos de las sub-líneas estudiadas (fig. 10), y representan alrededor del 30% de los trayectos. Las demás contienen en la mayoría de los casos uno o varios fragmentos compatibles con una de las sub-líneas (como se describe en la tabla 8d), aunque a veces (2% de las trayectorias) el estado de un vehículo no cambió entre las sub-líneas, por lo que una misma trayectoria contiene información relativa a más de un servicio.

3.1.4 Especificar los modelos de distribución de tiempos de viaje y de parada

Debido a sus ventajas computacionales, se han elegido dos familias de distribuciones normales (ecuación 14) para modelar los tiempos de viaje y de parada. Teniendo en cuenta los ciclos de movilidad de la ciudad, cada una de estas familias proporciona una función diferente para cada línea, parada, tipo de día (laborable, sábados, o domingos y festivos), periodo del año (verano o no), y clase temporal (con un lapso de 30min, y aproximadamente 16 horas diarias de servicio, hay 32 clases posibles: De 07:00 a 07:30, de 07:30 a 08:30, etc.).

$[p_{a,\tau,\gamma,\delta,\xi,\eta}$: t. viaje entre paradas con $p \in T; p \sim \mathcal{N}((\mu_p)_{a,\tau,\gamma,\delta,\xi}, (\sigma_p)_{a,\tau,\gamma,\delta,\xi}{}^2)]$,
$[u_{a,\tau,\gamma,\delta,\xi,\eta}$: t. parada con $u \in T;\ u \sim \mathcal{N}((\mu_u)_{a,\tau,\gamma,\delta,\xi}, (\sigma_u)_{a,\tau,\gamma,\delta,\xi}{}^2)]$, [$a$: id de la línea],
[τ: núm. de parada (para t. viaje, el de la parada inic.)], [γ: época del año, $\gamma \in$ {verano, resto del año}], [δ: tipo de día, $\delta \in$ {laboral, sábado, domingo o festivo}], [ξ: clase de h. del día, $\xi \in \{1 \ldots \eta\}$], [$\eta$: núm. de clases de h. $\eta \in \mathbb{N}$ (14)

Los tiempos de viaje y los tiempos de permanencia de los tramos de la ruta 1 se han caracterizado en cada parada por 192 distribuciones cada uno, según el periodo del año, el tipo de día, y la franja horaria aplicables. Sus medias y desviaciones estándar se han calculado utilizando las entradas pertinentes de la tabla avl_sintetizado.

3.1.5 Ensamblar los servicios

Tras aplicar el proceso descrito en el apartado 3.5, se encontraron 42319 posibles servicios.

3.1.6 Fusionar los casos en los que un vehículo ha cambiado su id a mitad de servicio

Este refinamiento permite detectar alrededor de 2 incidencias diarias de este problema, reduciendo el número de servicios candidatos a 41641.

3.1.7 Asignación de servicios a los viajes programados y actualización de los intervalos de tiempo de las visitas

40352 servicios han sido asignados a un inicio de servicio programado (111 utilizando un vehículo diferente al previsto); mientras que los otros 1289 no lo fueron.

3.1.8 Imputación de grupos de embarque

La aplicación de los criterios descritos en el apartado 2.8 proporciona los siguientes resultados:

- Se ha considerado que el 94,7% de los grupos de embarque informan correctamente de su ruta y parada de autobús.
- El 5% han sido asignados a otra parada que la registrada automáticamente.
- El 0,3% no estaba vinculado a ningún servicio. Es probable que representen casos en los que el estado del vehículo erróneamente reflejaba que viajaba por la línea 1.

3.1.9 Seleccionar los servicios respaldados por suficiente información

Tras considerar los resultados de los apartados 3.1.7 y 3.1.8, se han elegido los siguientes criterios de aceptación, utilizando la nomenclatura de la ecuación (13):

- Para los servicios asignados a un inicio programado (w = True):
    - Aceptar siempre si se ha utilizado el vehículo previsto (p = True).
    - Si se ha utilizado un autobús distinto al programado (p = Falso), exigir al menos 3 grupos de embarque vinculados al servicio ($h \geq 3$).
- Los servicios no programados requerirán más pruebas: al menos tres grupos de embarque y no menos de 12 entradas totales (un tercio del número de paradas de una sub-línea) que avalen su existencia ($h \geq 3 \wedge h + f \geq 12$).

Aplicando estos umbrales, la metodología reporta una media de 120 y 97 servicios diarios, según se analice un día laborable o no. En el primer caso, el 96,5% de los servicios habían sido previamente planificados, y se materializaron con el vehículo previsto; mientras que el 3% fueron planificados, pero ejecutados con un vehículo diferente; y el 0,5% fueron servicios no planificados. Durante los días no laborables, las ratios correspondientes son del 99,2%, 0,5% y 0,3%; lo que es coherente con el hecho de que los fines de semana y los días

festivos suelen ser menos exigentes para el transporte público de la ciudad, lo que se traduce en menos desviaciones del horario para reaccionar a la evolución del sistema de tráfico.

3.2 Análisis de los resultados

3.2.1 Caracterización de los servicios

Esta sección recoge varios ejemplos para ilustrar cómo esta metodología ha mejorado con éxito la caracterización de los servicios que estaban registrados en el IPTS de una manera que dificultaba su consideración.

3.2.1.1 Reconstrucción de un servicio a partir de información fragmentada y errónea

La figura 11 muestra el caso elegido para este análisis. El eje horizontal temporal se ha dividido en tres regiones con un desplazamiento entre ellas para facilitar su visualización:

- El central, donde se representa el servicio real detectado por la metodología y la salida prevista (5). Su eje temporal se ha colocado en la parte inferior del gráfico.
- La zona más a la izquierda, con su eje temporal situado en la parte superior de la figura. Incluye los datos brutos pertinentes de AVL y AFC, con un desplazamiento de -20min:
    - 4 identificadores de viaje AVL
        - (1): Desde “Arsenio Odriozola 16” hasta “San Fernando 66”, con un espacio de casi 1h entre “Plaza de Italia” y “Luis Martínez”.
        - (2): Desde “San Martín” hasta “Pctcan”, solapando con (1) a lo largo de sus primeras 9 paradas, faltando datos en “Avenida de Valdecilla” y “Torres Quevedo 22”.
        - (3): Un único evento, en la “Plaza de Italia”.
        - (4): Un único evento, en 'Pctcan', la última parada del servicio. Ocurre alrededor de medio minuto antes de que termine (2).
    - 19 eventos de AFC, que se producen entre la “Plaza de Italia” y “José Mª Cossío 24”.
- La zona de la derecha sólo contiene las horas de llegada y salida claramente no relacionadas registradas por el subsistema de planificación (6), con un desplazamiento de -40min.

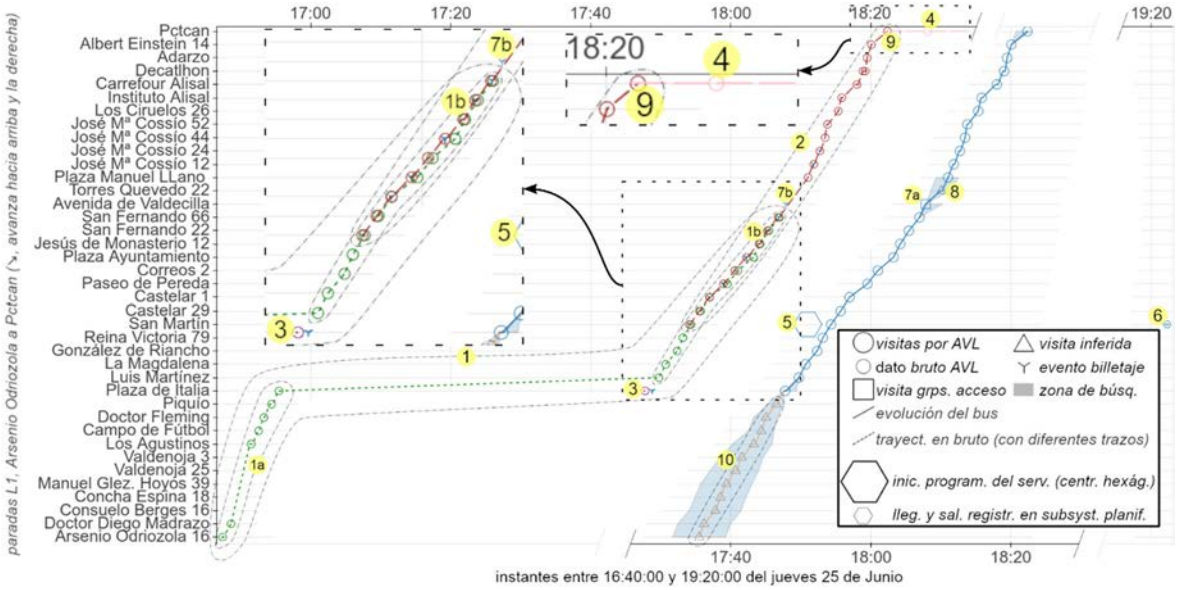


Figura 11: Caracterización de un servicio con información fragmentada y errónea

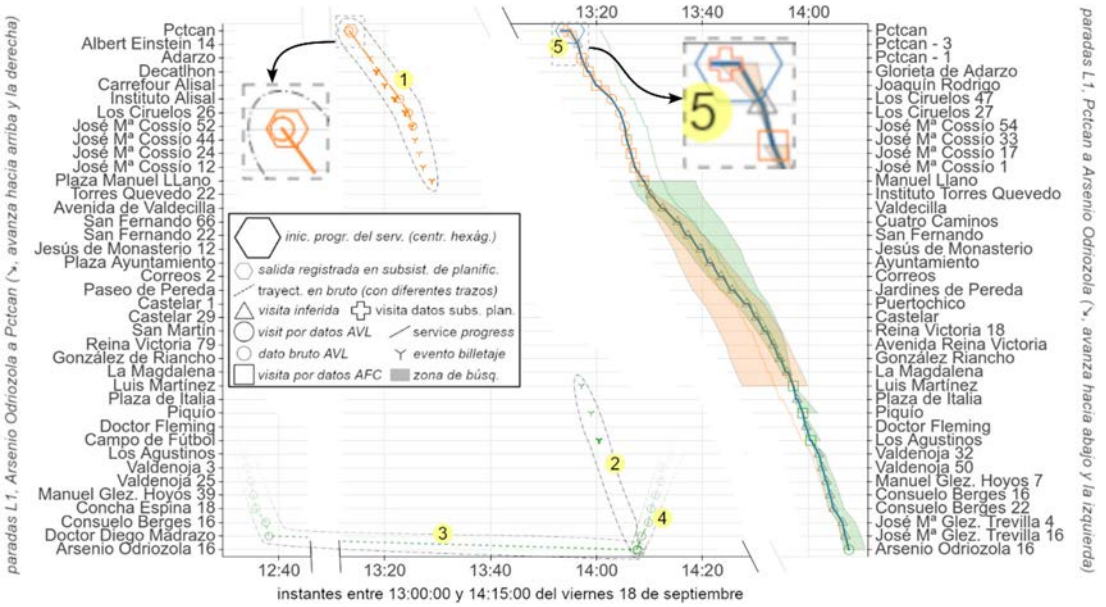


Figura 12: Caracterización de un servicio cuya id de vehículo cambia mientras sucede

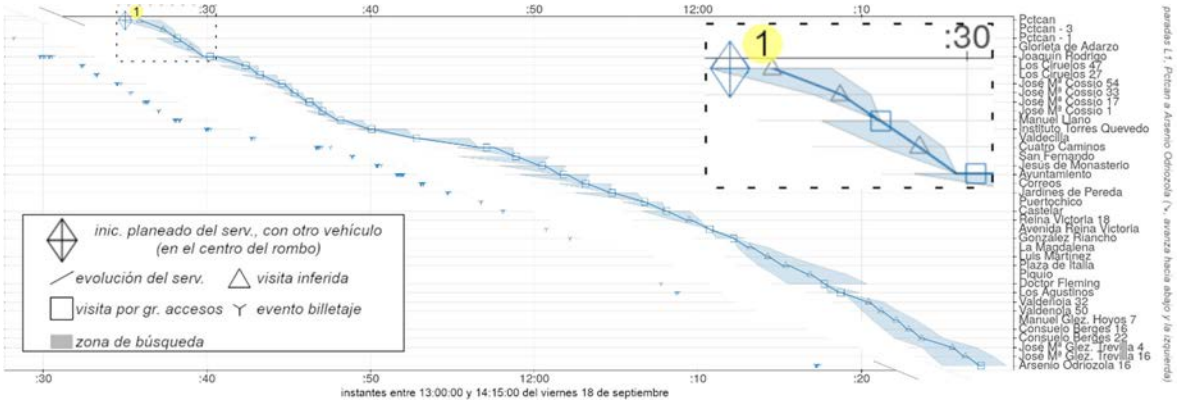


Figura 13: Servicio con sólo datos AFC. El vehículo utilizado no fue el previsto

El servicio propuesto se ha construido haciendo uso de la información disponible. La primera parte del viaje 1 se ha considerado como 2 fragmentos diferentes, descartando el primero (1a, que probablemente se debe a un estado incorrecto del vehículo) y utilizando el segundo (1b). Después de la última entrada de 1, la escala en "Avenida Valdecilla" (7a) se aproxima a partir de un evento de billetaje (7b); y la de "Torres Quevedo 22" (8) se infiere considerando las horas de salida y llegada de la parada anterior y siguiente, respectivamente. De las dos posibles llegadas a la terminal final (9), la del viaje 4, que ocurre 30s antes, es más probable según la hora de salida de "Albert Einstein 14" y la distribución de tiempos de viaje entre estas paradas durante el periodo de tiempo [17:30-18:00] en un día laboral.

Cabe destacar que, aunque el servicio estaba programado para comenzar en "San Martín", la metodología ha detectado con éxito que en realidad comenzaba unas paradas más arriba (en "Plaza de Italia", a partir del viaje 3). La búsqueda de eventos anteriores (10) no devolvió ninguna coincidencia.

3.2.1.2 Cambio de id de vehículo a mitad de servicio

La figura 12 muestra cómo aparece la información relativa a un servicio de la sub-línea de Pctcan a Arsenio Odriozola en el IPTS, y su caracterización por esta metodología. De nuevo, el eje temporal horizontal se ha dividido en tres zonas:

- La zona de la derecha, que contiene, con el eje temporal en la parte superior, los dos servicios detectados inicialmente, cómo se han combinado y el inicio previsto vinculado a ellos.
- Las regiones del centro y de la izquierda muestran, con desplazamientos de -40min y -20min y sus ejes temporales en la parte inferior, los registros brutos pertinentes.

Inicialmente, el paso 2.5 de la metodología había encontrado dos servicios:

- Uno para el vehículo 14 (naranja), respaldado por una trayectoria de 4 paradas, y varios eventos de billetaje (1), siendo el último en "Manuel Llano".
- Otro para el vehículo 224 (verde), deducido de 4 eventos de billetaje en 3 paradas (2 , el más temprano en "Luis Martínez"), y cualquiera de los dos eventos en bruto de AVL con la misma marca temporal en la terminal "Arsenio Odriozola 16", que forman parte de trayectorias opuestas que terminan (3) o comienzan (4) en ese lugar.

Se ha detectado, como se describe en el apartado 2.6, que estos servicios están formados con información parcial de uno que los engloba (se muestra con una línea azul más gruesa). Su correspondiente entrada en el subsistema de programación (5) sólo ha detectado la salida del vehículo, un poco más tarde que los datos de AVL disponibles en esa parada. Como está dentro del rango de viabilidad de "Pctcan – 1", se acepta y se utiliza para actualizar la hora de salida en "Pctcan", y para mejorar la llamada inferida en la parada intermedia "Pctcan – 3".

3.2.1.3 No hay datos de AVL y la identificación del vehículo es incorrecta

La figura 13 muestra un caso que ilustra dos situaciones que se dan en el caso de uso: que el subsistema AVL no registre ninguna entrada y que un vehículo diferente al previsto realice el servicio.

Hay un desplazamiento de 10 minutos entre el lugar donde se dibuja el servicio y la salida programada (parte derecha, eje temporal en la parte superior), y donde se encuentran los datos brutos del AFC (a la izquierda, eje temporal en la parte inferior). Se puede observar (1) que, como el subsistema de planificación no registró el inicio del servicio, las escalas en "Pctcan" y "Pctcan - 3" tuvieron que inferirse utilizando la llegada a "Pctcan - 1" como punto fijo.

3.2.2 Tratamiento de las terminales donde se inician los servicios

El objetivo de esta sección es estudiar el beneficio de la forma en que esta metodología maneja los datos disponibles en los terminales especialmente problemáticos, como ocurre en esta ruta. Para ello, los 25466 servicios que presentan tiempos de salida registrados del subsistema de planificación que, como se describe en el apartado 2.7, han sido aceptados para su caracterización, se utilizarán como verdad de base para compararlos con los resultados obtenidos en tres escenarios en los que no se tendrá en cuenta esa información:

(A): Sigue el comportamiento de la metodología por defecto para una ruta cuando el subsistema de programación no registró el inicio de un servicio.

(B): Si los datos de la primera parada se consideran creíbles, utilícelos de la misma manera que en cualquier otra parada.

(C): Si el inicio planificado de un servicio cae dentro de su rango de búsqueda (tabla 10c) correspondiente (ya almacenado, o calculado utilizando la escala más cercana del servicio sustentada en datos reales), se utilizará como salida, si ocurre más tarde que cualquier entrada AFC o AVL disponible. Esto equivale a suponer que el cumplimiento del horario es lo suficientemente estricto como para confiar en las horas de salida previstas, a menos que sean imposibles o muy improbables.

La figura 14 muestra las distribuciones del error absoluto de la hora de salida del servicio comunicada en los escenarios A y C. Los servicios se han clasificado en función de su “espacio vacante”: a qué distancia (medida en tramos de viaje) se encuentran sus primeras visitas basadas en los datos de AVL o AFC respecto a sus inicios programados. Como puede observarse, la decisión de basarse en la hora de inicio inferida en lugar de la prevista proporciona aproximaciones con menor dispersión (desviaciones estándar de 13s y 17s, respectivamente) y un menor error medio absoluto (MAE), aunque a medida que aumenta la incertidumbre (más llamadas desconocidas entre el inicio del servicio y el primer punto de datos) esta ventaja disminuye.

Los escenarios A y B sólo difieren para aquellos servicios en los que se pueden encontrar datos AVL o AFC compatibles en la primera parada programada (espacio vacante cero). La figura 15 muestra sus distribuciones de errores absolutos en este caso. De nuevo, el escenario A infiere los datos que faltan con menos dispersión (desvíos estándar de 15s y 17s, respectivamente) y MAE (11s frente a 13s).

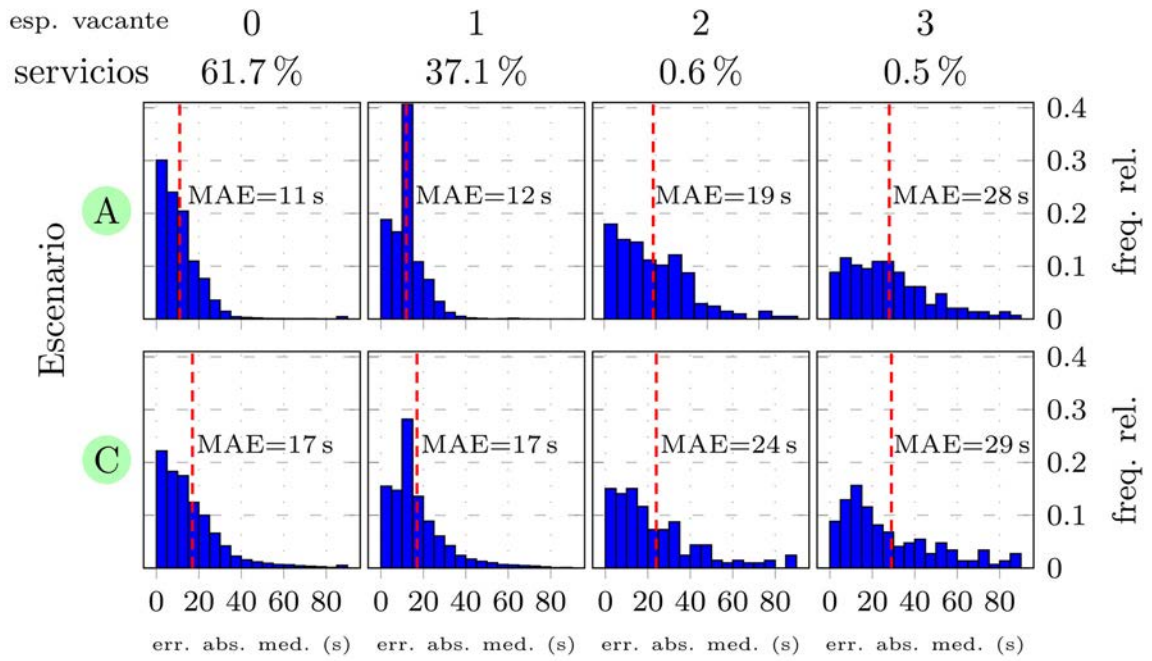


Figura 14: Distr. de errores abs. de hora de salida para los escenarios A y C

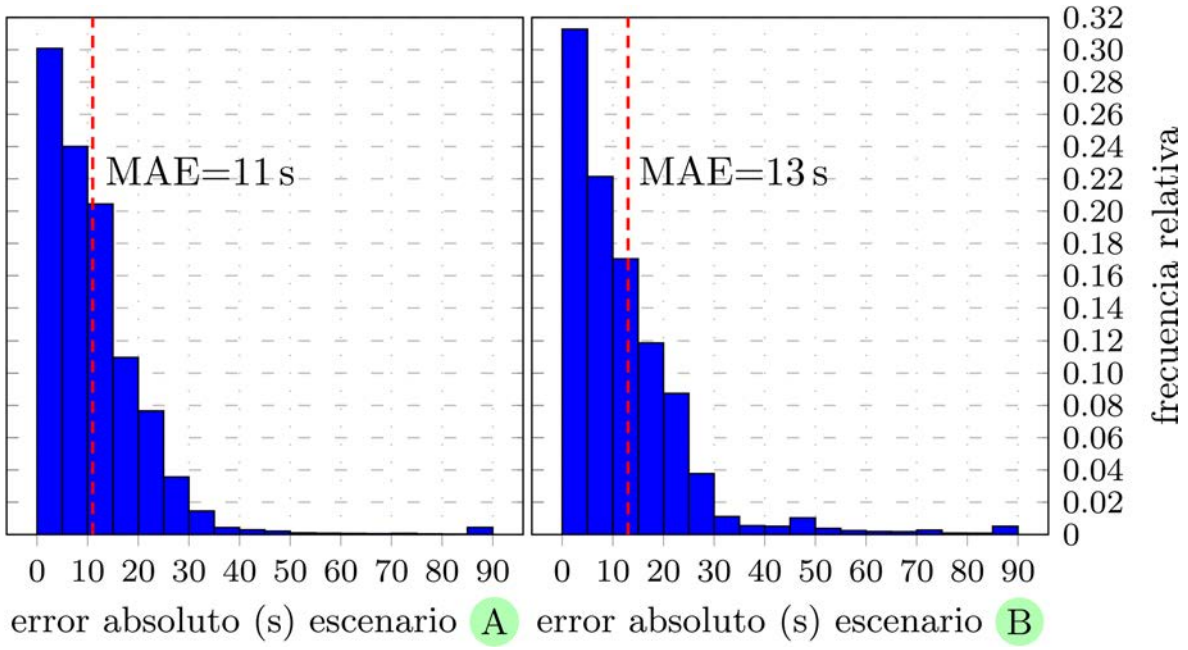


Figura 15: Distr. de los errores abs. de la hora de salida en los escenarios A y B

3.2.3 Solidez frente a los datos ausentes y erróneos

En esta sección se analiza cómo afecta a la metodología información de AVL y de detección del inicio del servicio incompleta y errónea (los eventos de emisión de billetes están totalmente disponibles en todos los escenarios). Los 16863 servicios en los que los subsistemas de planificación y AVL registraron todas las escalas (el 49% del total) se utilizarán como datos reales; y se compararán con los resultados de ejecutar esta metodología utilizando sólo una parte de los datos AVL brutos registrados y de las detecciones del subsistema de planificación, elegidos mediante un muestreo Bernoulli; añadiendo también diferentes cantidades de lecturas erróneas AVL sintéticas, que se han generado aleatoriamente siguiendo estas reglas:

- La *parada*, el *vehículo* y la *UID de grupo* se eligen entre todos sus valores presentes.
- *Instante* ocurre entre las 07:00 y las 23:00 de cualquier día del año.
- El muestreo de la distribución de duraciones se simula utilizando sus percentiles y la Distribución Uniforme.

En la fig. 16, los porcentajes son relativos a las lecturas de AVL en bruto y a los servicios planificados disponibles en el conjunto de datos. Por ejemplo, un escenario con un 25% de datos reales y un 75% de errores simulados sólo lee la llegada y la salida de los vehículos en la parada inicial registrada por el subsistema de programación en el 25% de los servicios programados; mientras que sus registros de AVL en bruto se crea combinando un muestreo de Bernoulli de la información real con una probabilidad del 25% y una cantidad tres veces superior de entradas falsas.

A medida que se dispone de más datos reales en un escenario, los servicios se caracterizan con mayor precisión. Por ejemplo, con una muestra relativamente pequeña (25%), si bien el percentil 99 no difiere significativamente de no utilizar AVL o de detectar el inicio del servicio en absoluto (algo menos de 7min), ya se puede apreciar que el MAE es bastante más probable que sea menor: el cuartil inferior, la mediana y el cuartil superior se reducen de 4s, 9s y 24s a 0s, 4s y 13s, respectivamente (A y B).

También es notable la resistencia de la metodología contra las entradas incorrectas artificiales, que crece a medida que se dispone de más lecturas verdaderas en el escenario. Dos ejemplos son:

- Con sólo el 25% de los datos reales, añadir cuatro veces más entradas erróneas sólo aumenta el cuartil superior de 6m50s a 7m7s (B y C).
- Si toda la información real está disponible, la metodología identifica con éxito los valores correctos como semillas, y es capaz de ignorar por completo muchos eventos falsos (D y E).

4. CONCLUSIÓN

La metodología descrita en este trabajo es capaz de combinar la información del AFC, del AVL y del subsistema de programación para proporcionar una mejor caracterización de los servicios de las rutas ofrecidas en un Sistema de Transporte Público; mejorando los problemas que comúnmente ocurren cuando se trabaja con datos del IPTS. Los eventos cuyos atributos los clasifican erróneamente como parte de diferentes servicios son identificados y tratados adecuadamente, así como aquellos que probablemente no hayan ocurrido realmente. Las escalas en cada parada de cada servicio se delimitan combinando las múltiples fuentes de datos disponibles en ese caso concreto, proporcionando las horas de llegada y salida más probables si no las hay. Asimismo, se formula una forma de detectar y tratar aquellos casos en los que un vehículo cambia su identificación a mitad de servicio, lo que llevaría a una tergiversación de su perfil de carga.

Se ha presentado un estudio de caso, en el que varios ejemplos ilustran algunos de los problemas que resuelve esta metodología (información fragmentada y errónea, cambio de identificación del vehículo a mitad de servicio y caracterización de un servicio no planificado utilizando únicamente datos AFC). También se utiliza para analizar la repercusión de la forma en que se tratan los datos de los terminales y cómo se comporta la metodología con diferentes proporciones de información buena y errónea.

Como próximo objetivo, los autores están trabajando actualmente en la aplicación de la metodología de cadenas de viaje con los eventos de billetaje y los servicios caracterizados por esta metodología, para proporcionar perfiles de carga de vehículos y matrices OD más precisos. Otras posibles líneas de investigación son la utilización de otras distribuciones para modelar los tiempos de permanencia y de viaje, o la aplicación de modelos más detallados para estimar a partir de los eventos de billetaje la llegada y la salida en una parada cuando no se dispone de registros AVL.

REFERENCIAS

AIMSUN (2020). Líneas de transporte público. En Aimsun Next 8.4 Manual del usuario. Aimsun.

ALSGER, A. A. M. Y ENG, B. (2016). Estimación de las matrices de origen y destino del tránsito utilizando datos de tarifas de tarjetas inteligentes Escuela de Ingeniería Civil. Informe técnico.

BERTINI, R. L. Y EL-GENEIDY, A. (2003). Generating transit performance measures with archived data. Transportation Research Record, (1841):109-119.

BUNEMAN, K. (1984). Automated and Passenger-Based Transit Performance Measures. Transportation Research Record, p. 23-28.

CHEN, B. Y., SHI, C., ZHANG, J., LAM, W. H., LI, Q. Y XIANG, S. (2017). Algoritmo de búsqueda de rutas más fiable para maximizar la probabilidad de llegada a tiempo. Transportmetrica B, 5(3), p. 253-269.

CHEPURI, A., RAMAKRISHNAN, J., ARKATKAR, S., JOSHI, G., Y PULUGURTHA, S. S. (2018). Examen de los indicadores de rendimiento basados en la fiabilidad del tiempo de viaje para las rutas de autobús utilizando datos de trayectoria de autobuses basados en GPS en la India. Revista de Ingeniería de Transporte Parte A: Sistemas, 144(5).

CHU, K. K. A., CHAPLEAU, R., Y TRÉPANIER, M. (2009). Encuesta sobre los viajes en tránsito pasivo de los conductores con un sistema de cobro automático con tarjeta inteligente y aplicaciones. Transportation Research Record, 2105, p.1-10.

DAI, Z., MA, X., Y CHEN, X. (2019). Modelización del tiempo de viaje en autobús utilizando datos de la sonda GPS y de la tarjeta inteligente: Un enfoque probabilístico que considera el tiempo de viaje del enlace y el tiempo de permanencia en la estación. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations, 23(2), p. 175-190.

DE PABLOS HEREDERO, C., PÉREZ BERMEJO, L. J., Y MONTES BOTELLA, J. L. (2012). Impacto de los sistemas de apoyo a la explotación (SAE) en la mejora de los servicios de transporte público urbano. Cuadernos de Economía y Dirección de la Empresa, 15(1), p.12-24.

FURTH, P. G., HEMILY, B. J., MULLER, T. H. J., Y STRATHMAN, J. G. (2003). Documento Web 23 del TCRP (Proyecto H-28): Contractor's Final Report Uses of Archived AVL-APC Data to Improve Transit Performance and Management: Review and Potential. 23 (junio de 2003).

GOKASAR, I. Y CETINEL, Y. (2019). Una nueva estrategia para el diagnóstico de las cabeceras de los autobuses utilizando datos avl. MT-ITS 2019 - 6ª Conferencia internacional sobre modelos y tecnologías para sistemas de transporte inteligentes.

GOOGLE. Visión general de GTFS Static - Static Transit - Google Developers.

GSCHWENDER, A., MUNIZAGA, M., Y SIMONETTI, C. (2016). Uso de datos de tarjetas inteligentes y GPS para la política y la planificación: El caso de Transantiago. Investigación en Economía del Transporte, 59, p. 242-249.

HARSHA, M., MULANGI, R. H., Y KUMAR, H. D. (2020). Análisis de la variabilidad del tiempo de viaje de los autobuses utilizando datos de localización automática de vehículos. Transportation Research Procedia, 48(2018), p. 3283-3298.

HOUNSELL, N. Y SHRESTHA, B. (2012). Un nuevo enfoque para la prioridad cooperativa de los autobuses en los semáforos. IEEE Transactions on Intelligent Transportation Systems, 13(1), p. 6-14.

HUANG, Z., XU, L., LIN, Y., WU, P., Y FENG, B. (2019). Identificación del comportamiento de transferencia de metro a autobús en toda la ciudad basada en datos combinados de tarjetas inteligentes y GPS. ciencias aplicadas.

JIANG, X. Y YANG, X. (2014). Modelos basados en la regresión para el tiempo de permanencia en el autobús. 2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014, p. 2858-2863.

KOSHY, R. Z. Y ARASAN, V. T. (2005). Influencia de las paradas de autobús en las características de flujo del tráfico mixto. Journal of Transportation Engineering, 131(8), p. 640-643.

LI, M., Zhou, X., y Rouphail, N. M. (2017). Cuantificación de la variabilidad del tiempo de viaje en un solo cuello de botella basado en distribuciones estocásticas de capacidad y demanda. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations, 21(2), p. 79-93.

LIN, Y., YANG, X., ZOU, N., Y JIA, L. (2013). Predicción del tiempo de llegada de los autobuses en tiempo real: Estudio de caso para Jinan, China. Journal of Transportation Engineering, 139(11), p. 1133-1140.

LUO, D., BONNETAIN, L., CATS, O., Y VAN LINT, H. (2018). Construyendo perfiles de carga espacio-temporales de vehículos de tránsito con múltiples fuentes de datos. Registro de investigación de transporte, 2672(8), p. 175-186.

MA, X., PH, D., ASCE, M., WANG, Y., PH, D., Y ASCE, M. (2014). Desarrollo de una plataforma impulsada por datos para las medidas de rendimiento de tránsito utilizando datos de tarjetas inteligentes y GPS. Journal of Transportation Engineering, 140(12), p. 1-13.

MAPZEN. FUNDACIÓN E INTERLINE TECHNOLOGIES LLC. Transitland. MobilityData. Especificación general de alimentación de tránsito.

MOBILITYDATA IO. OpenMobilityData - Fuentes de transporte público de todo el mundo.

MOREIRA-MATIAS, L. (2016). Hacia un modelo de estimación de la demanda basado en AVL.

GRUPO PTV (2020). Uso de las distribuciones de tiempo. En el manual de usuario de PTV Vissum 2021, página 240.

QU, X., OH, E., WENG, J., Y JIN, S. (2014). Análisis de la fiabilidad del tiempo de viaje en autobús: Un estudio de caso. Actas de la Institución de Ingenieros Civiles: Transport, 167(3):178-184.

RAJBHANDARI, R., CHIEN, S. I., Y DANIEL, J. R. (2003). Estimación de los tiempos de permanencia de los autobuses con información del contador automático de pasajeros. Transportation Research Record, (1841):120-127.

RASHIDI, S. Y RANJITKAR, P. (2013). Aproximación y predicción a corto plazo del tiempo de permanencia de los autobuses utilizando datos de AVL. Journal of the Eastern Asia Society for Transportation Studies, 10:1281-1291.

SAGHAEI, H. (2016). Design and Implementation of a Fleet Management System Using Novel GPS/GLONASS Tracker and Web-Based Software-Automatic Vehicle Locator (AVL), Fleet Management System (FMS), Global Positioning System (GPS), Global Navigation Satellite System (GLONASS), Ge.

SPIEGEL. Santander: Un prototipo de ciudad digital inteligente en España - spiegel online.

SRINIVASAN, K. K., PRAKASH, A. A., Y SESHADRI, R. (2014). Finding most reliable paths on networks with correlated and shifted log-normal travel times. Transportation Research Part B: Methodological, 66:110-128.

TAYLOR, M. Y SUSILAWATI (2012). Modelización de la fiabilidad del tiempo de viaje con la distribución de Burr. Procedia - Social and Behavioral Sciences, 54(Noviembre 2014):75-83.

TRÉPANIER, M., MORENCY, C., Y AGARD, B. (2009). Calculation of Transit Performance Measures Using Smartcard Data. Journal of Public Transportation, 12(1):79-96.

TRÉPANIER, M., TRANCHANT, N., Y CHAPLEAU, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. Journal of Intelligent Transportation Systems: Technology, Planning, and Operations, 11(1):1-14.

VAN DIGGELEN, FRANK, ENGE, P. (2015). El primer MOOC de GPS del mundo y el laboratorio mundial que utiliza teléfonos inteligentes. En 28th International Technical Meeting of the Satellite Division of The Institute of Navigation, páginas 361-369, Tampa, Florida.

WIKIMEDIA. File:Spain map modern.png - Wikimedia Commons.

WILSON, N. H., ZHAO, J., Y RAHBEE, A. (2009). The potential impact of automated data collection systems on urban public transport planning. Operations Research/ Computer Science Interfaces Series, 46(1):75-99.

ZHANG, H., CUI, H., Y SHI, B. (2019). Un análisis basado en datos para el rendimiento operativo de los vehículos de la red de transporte público. IEEE Access, 7:96404-96413.