UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA

Programa de Doctorado en Empresa, Internet y Tecnologías de las Comunicaciones (EmITIC)

Tesis por compendio de publicaciones

"Modelización del análisis de la demanda y de la calidad de servicio en el transporte público regular de viajeros por carretera mediante Minería de Datos"

Teresa Cristóbal Betancor

Las Palmas de Gran Canaria

Abril 2019

UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

D. Miguel Ángel Ferrer Ballester COORDINADOR DELPROGRAMA DE DOCTORADO Empresa, Internet y Tecnologías de la Información y Comunicaciones DE LA UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA,

INFORMA,

Que la Comisión Académica del Programa de Doctorado,

en su sesión de fecha 12 de abril de 2019, tomó el acuerdo de dar el consentimiento para su tramitación, a la tesis doctoral titulada " Modelización del análisis de la demanda y de la calidad de servicio en el transporte público regular de viajeros por carretera mediante Minería de Datos" presentada por la doctoranda Dª Teresa Cristóbal Betancor y dirigida por los Doctores Dr. D. Carmelo Rubén García Rodríguez y D. Alexis Quesada Arencibia.

Y para que así conste, y a efectos de lo previsto en el Artº 11 del Reglamento de Estudios de Doctorado (BOULPGC 7/10/2016) de la Universidad de Las Palmas de Gran Canaria, firmo la presente en Las Palmas de Gran Canaria, a doce de abril de dos mil diecinueve.

UNIVERSIDAD DE LAS PALMAS
DE GRAN CANARIA

Tesis por compendio de publicaciones

Título	Modelización del análisis de la demanda y de la calidad de servicio en el transporte público regular de viajeros por carretera mediante Minería de Datos
Programa de Doctorado	Empresa, Internet y Tecnologías de las Comunicaciones (EmITIC)
Unidad Responsable	Escuela de Doctorado de La Universidad de Las Palmas de Gran Canaria (EDULPGC)
Lugar	Las Palmas de Gran Canaria
Fecha	Abril de 2019

Directores

Dr. Carmelo R. García Rodríguez **Dr. Alexis Quesada Arencibia**

Autora

Teresa Cristóbal Betancor

A mis padres, y a Rubén, Julia y Pablo.

Agradecimientos

Quiero agradecer a mis directores de tesis Dr. Alexis Quesada y Dr. Carmelo Rubén García, y a Dr. José Javier Lorenzo su dedicación y consejo, a la empresa de transporte Global S.U. que cediera sus datos para la realización de estos trabajos, y reconocer la generosidad de todos aquellos que con su ayuda los hicieron posible.

Índice

1. Introducción

Los medios de transporte juegan un papel fundamental en la sociedad actual, permitiendo la circulación de bienes y de personas, favoreciendo el comercio, el desarrollo económico y el desarrollo humano. El filósofo y economista Amartya Sen, galardonado con el Premio en Ciencias Económicas en memoria de Alfred Nobel en 1998, en su libro "Development as Freedom" indicó que "*... el Desarrollo consiste en remover las barreras a la libertad que dejan a la gente con pocas opciones y limitadas oportunidades de ejercer su capacidad de decidir. ... Desde una perspectiva opuesta, la imposibilidad de moverse por el territorio roba a las personas la libertad de satisfacer muchas de sus necesidades como la educación, la salud, la alimentación, el ocio y el contacto con seres queridos, entre otras" [1].*

El auge a lo largo de la segunda mitad del siglo XX de un modelo de transporte por carretera basado en el coche particular ha provocado serios inconvenientes que, entre otras consecuencias, afectan directamente a la vida humana: más de 3.700 personas mueren en las carreteras del mundo cada día, decenas de millones resultan heridas o discapacitadas cada año [2], y 4.2 millones mueren a causa de la contaminación medioambiental generada por los vehículos [3].

Con la voluntad de encontrar alternativas a este modelo y reorientar la movilidad hacia un enfoque más sostenible se hace imprescindible, entre otras cuestiones, fomentar el uso del transporte público frente al privado y crear una cultura de mejora continua de los sistemas de transporte basada en la recolección y preservación de datos, que facilite la comprensión y gestión del tráfico y de la demanda, escuchando y profundizando en las necesidades de la ciudadanía.

En el ámbito concreto del transporte público por carretera, las empresas operadoras y las autoridades responsables necesitan información para configurar debidamente la red de transporte; para planificar, ejecutar y monitorizar las expediciones; para administrar la flota de vehículos, controlar costos y gastos, y gestionar medios de pago; y para medir la calidad y la satisfacción de los usuarios, conocer sus hábitos y comportamientos tanto de forma individual como colectiva. Tal es así que la información puede considerarse un recurso necesario para la gestión del servicio de transporte y parte integradora del servicio de movilidad en sí mismo [4].

El conjunto de avances tecnológicos de los que se ha beneficiado este sector, especialmente en lo que se refiere a las comunicaciones y a la informática, a las Tecnologías de la Información, ha dado lugar a un campo de estudio denominado

“Sistemas de Transporte Inteligentes” (STI), en el que el uso combinado de los sensores, comunicaciones móviles y los sistemas de cómputo permiten compartir la información entre plataformas para desarrollar una visión más amplia de la operativa [5]. Dada la importancia que para el sector tienen estos sistemas y con la finalidad de fomentar su uso coordinado y coherente, las autoridades han promovido el desarrollo de recomendaciones y estándares que se han plasmado en la Unión Europea en la *Directiva 2010/40/UE del Parlamento Europeo y del Consejo de 7 de julio de 2010 por el que se establece el marco de implantación de los sistemas de transporte inteligentes en el sector del transporte por carretera y las interfaces con otros modos de transporte*, donde entre otras cuestiones se puntualiza:

> *”Los sistemas de transporte inteligentes son aplicaciones avanzadas que proporcionan servicios innovadores en relación con los diferentes modos de transporte y la gestión del tráfico y permiten a los distintos usuarios estar mejor informados y hacer un uso más seguro, más coordinado y «más inteligente» de las redes de transporte. ... Integran las telecomunicaciones, la electrónica y las tecnologías de la información con la ingeniería de transporte con vistas a planear, diseñar, manejar, mantener y gestionar los sistemas de transporte”*

En el ámbito de esta directiva, en concreto atendiendo a la Acción Prioritaria A, se ha desarrollado una normativa europea de modelo de datos para el transporte público “Transmodel”, norma UNE-EN 12896, que proporciona un marco para los modelos de datos de toda el área de operaciones del transporte público, con objeto de hacer posible que los operadores, las autoridades y los proveedores de software trabajen juntos hacia sistemas integrados, y para garantizar que los desarrollos futuros puedan adaptarse sin dificultad [6]. De los términos definidos en este modelo formal, son de especial interés:

- **Red de transporte**: conjunto de entidades que representan las vías, puntos de interés y rutas, que describen el espacio donde se desarrollan las operaciones de los vehículos.

- **Parada**: punto de la red de transporte donde los pasajeros embarcan o desembarcan de los vehículos.

- **Ruta**: recorrido que se realiza de manera sistemática, que comienza y termina en una parada de la red, y que pasa de manera ordenada por un conjunto de paradas.

- **Planificación del transporte**: conjunto de actividades a las que se le asignan recursos (un vehículo y un conductor) y que han de ser ejecutadas un día y hora determinadas. De entre ellas, una de especial relevancia es la actividad denominada “servicio de línea” o “expedición”, que es la que realiza un vehículo al recorrer una ruta determinada, el día y la hora previamente planificada.

- **Horas de paso programadas**: conjunto ordenado de tiempos de paso por las paradas de una ruta de un determinado servicio de línea, en cuyo cálculo hay que considerar tanto el tiempo que tarda el vehículo en realizar el trayecto entre paradas consecutivas como el tiempo que tardan los pasajeros en subir y bajar de los vehículos.

- **Base de datos de transporte**: base de datos que contiene todas las entidades y sus relaciones, los recursos asignados y los datos registrados en las operaciones de transporte, que han podido ser generados por los sistemas embarcados en los vehículos.

Por último mencionar otro modelo europeo que responde a la problemática relacionada con la evaluación de la calidad en el servicio de transporte público, la norma UNE-EN 13816, y que determina los distintos ámbitos en los que han de desarrollarse sistemas de gestión de la calidad, a los que denomina: Servicio ofertado, Accesibilidad, Información, Tiempo, Atención al cliente, Seguridad, Confort e Impacto.

En los artículos presentados en este documento se propone el uso de metodologías y técnicas de ciencia de datos con el fin de descubrir conocimiento en dos de los aspectos fundamentales en el sector del transporte público de viajeros por carretera: la demanda y los tiempos de viaje. En ellos se plantean modelos basados en arquitecturas STI y en el modelo de datos estándar de la Unión Europea Transmodel, a partir de los datos históricos registrados en dos de los sistemas existentes de manera generalizada en los vehículos, en los sistemas de posicionamiento y en los de medio de pago, y almacenados en la base de datos de transporte. El objetivo principal de todos ellos es generar nueva información para medir la calidad en el ámbito “Servicio ofertado” y en el de “Tiempo”, concretamente evaluando la adecuación a las necesidades del cliente a partir de la demanda, el cumplimiento horario y la duración de los viajes.

El descubrimiento de conocimiento es la extracción no trivial de información de datos implícita, previamente desconocida y potencialmente útil. Dado un conjunto de hechos (datos) *F*, un lenguaje *L* y cierta medida de certeza *C*, definimos un patrón como una declaración *S* en *L* que describe las relaciones entre un subconjunto F_S de *F* con una certeza *c*, de modo que *S* es más simple (en algún sentido) que la enumeración de todos los hechos en F_S. Un patrón que es interesante (de acuerdo con una medida de interés impuesta por el usuario) y suficientemente cierto (de nuevo de acuerdo con los criterios del usuario) se llama conocimiento [7]. En el descubrimiento de conocimiento en las bases de datos intervienen distintas disciplinas como sistemas expertos, estadística, bases de datos, visualización de datos, aprendizaje automático, computación de alto rendimiento, etc. A la aplicación de algoritmos estadísticos y de aprendizaje automático para la extracción de patrones o modelos de los datos, se le denomina Minería de Datos [8].

Con la estadística inductiva es posible, a partir de una muestra, dar respuestas a preguntas formuladas como hipótesis, estimar numéricamente características, correlaciones o modelar relaciones entre variables, pero no siempre dan buenos resultados con datos multidimensionales complejos [9]. Por otro lado, las técnicas de aprendizaje automático combinan elementos de aprendizaje, adaptación, evolución y lógica difusa para crear modelos “inteligentes”, en el sentido de que la estructura emerge a partir de unos datos no estructurados, siendo por esta razón la disciplina en la que se sustenta el desarrollo de los trabajos presentados en esta tesis.

1.1. Estado del arte

Para poder alcanzar los objetivos de eficiencia y calidad de servicio en el transporte público, un requisito fundamental es conocer las necesidades y hábitos de movilidad de las personas. A partir de este conocimiento, se pueden realizar con garantías los tres procesos básicos en los que se basa esta actividad que son, diseño de la red de transporte, planificación de servicios y control de operaciones. En [10] se realiza una exhaustiva revisión de las técnicas utilizadas en el diseño de la red de transporte y en la planificación de las operaciones, y en el lado del análisis de la calidad en [11] se realiza una revisión exhaustiva de las técnicas utilizadas para analizar el comportamiento y evaluar los principales parámetros que le afectan.

A continuación, se presenta una revisión bibliográfica de trabajos en los que se ha utilizado técnicas y métodos de Minería de Datos para resolver algunos de los problemas presentes en los sistemas de transporte. En función de la fuente de datos utilizada, estos trabajos se pueden clasificar en dos grupos: los que se basan en datos relacionados con los movimientos de los viajeros y los que utilizan datos relacionados con la posición de los vehículos en la red de transporte, y en ambos se encuentran publicaciones que abordan los tres principales problemas que hay que afrontar para conseguir eficiencia y calidad de servicio: el diseño de la red de transporte, la planificación de servicios y el control de operaciones.

Trabajos relacionados con los datos de los movimientos de los viajeros

Entre los trabajos que utilizan datos asociados a los viajes realizados por los usuarios destacan los que, partiendo de los registros que se generan con el uso de las tarjetas de pago inteligentes, tienen por finalidad obtener conocimiento acerca de los perfiles y hábitos de uso de los usuarios [12], medir el uso de las infraestructuras de la red por parte de los viajeros [13], o realizar predicciones acerca de los tiempos de viaje para desarrollar servicios de información personalizada para el viajero [14], [15]. En [16] los autores además incorporan factores de carácter socio-demográfico: ubicación de centros comerciales, zonas deportivas, residenciales, etc. Atendiendo al análisis realizado en [9], los trabajos que proponen técnicas para obtener los patrones de movilidad de los usuarios en los sistemas de tránsito masivo se agrupan en dos categorías: los basados en métodos estadísticos, capaces de suministrar un modelo auto explicativo como resultado de un proceso estocástico, y los que utilizan redes neuronales. A partir de series temporales constituidas por los viajes realizados en ciertos intervalos de tiempo para predecir la demanda, en [17] se propone el uso de modelos estadísticos, en [18] se utilizan redes neuronales, y como ejemplo de procedimientos mixtos en [19] se introduce un proceso de selección de las funciones generadas antes de aplicar la red neuronal. En [20] se analiza el resultado de dos modelos diferentes de redes usando características temporales de la demanda observada (tendencia, ciclo y periodicidad), y en [21] se desarrolla un nuevo algoritmo de optimización híbrida, con técnicas de teorías de conjuntos y redes neuronales, para predecir el volumen de pasajeros por carretera. Como ejemplo de otras técnicas, en [22] se estudia el comportamiento espacial y temporal de los viajeros en la red de metro partiendo del uso de las tarjetas, utilizando técnicas de agrupamiento.

Trabajos relacionados con los datos de posicionamiento de los vehículos

Los datos de posicionamiento de los vehículos de transporte público se han utilizado fundamentalmente para mejorar el diseño de la red de transporte, para evaluar la calidad de servicio y también para realizar predicciones del tiempo de viaje, y como ejemplo de las distintas cuestiones abordadas con estos datos se presentan las siguientes referencias. En [23] se propone una metodología para evaluar la red de carretera analizando los tiempos de viaje mediante funciones de distribución estadísticas. En [24], mediante técnicas de agrupamiento desarrolladas por los propios autores, se analiza el impacto de la demanda y del tráfico en el rendimiento de las operaciones. Considerando la subida y bajada de los pasajeros a los vehículos, en [25] se realiza un estudio para evitar el hacinamiento que, junto con los retrasos en las llegadas, puede disuadir al ciudadano de utilizar los servicios públicos de transporte, y en [26] se generan diagramas de diagnóstico de la fiabilidad del servicio para descubrir cómo afecta la variabilidad de los atributos de los servicios en el comportamiento de los viajeros. En [27] se propone una metodología para la mejora del diseño de la red de transporte que incluye: detección y clasificación de las paradas, generación de los recorridos y estimación los tiempos de paso por las paradas, a partir de los datos GPS de los vehículos utilizando técnicas de agrupamiento. En [28] se propone una nueva métrica para evaluar la puntualidad de los autobuses utilizando los datos de posicionamiento de los vehículos. En [29] analizan las causas que originan irregularidades en la planificación de servicios. En el contexto de los sistemas de tránsito masivo por carretera planificados por horas de paso, utilizados fundamentalmente en redes interurbanas, en [30], [31] utilizan técnicas de agrupamiento y métricas ad-hocs para procesar los datos de posicionamiento de los vehículos y de movimiento de pasajeros y conseguir el mejor agrupamiento que permita evaluar la calidad de servicio proporcionado.

En lo que se refiere a la predicción del tiempo de viaje procesando únicamente datos de posicionamiento mediante técnicas de aprendizaje automático, existe un amplio conjunto de trabajos. En [32] se utilizan redes neuronales, técnicas de clasificación en [33] y de agrupamiento en [34]. También existe un número considerable de propuestas que abordan la predicción del tiempo de viaje empleando modelos de estado [35] y series temporales [36][37].

Hay que hacer notar que la mayoría de estos trabajos se han desarrollado en el contexto del transporte público urbano. Los estudios específicos sobre el transporte público interurbano por carretera no han recibido tanta atención, ni en lo que se refiere al análisis de la demanda ni a la evaluación sistemática de la calidad del servicio. En el caso del transporte público urbano, el objetivo de la planificación de servicios es determinar una frecuencia de paso de los vehículos por las paradas que garantice un nivel de calidad de servicio adecuado. Por el contrario, el objetivo en el caso del transporte interurbano es fijar unos horarios de paso que garanticen esta calidad y, como consecuencia de esta diferencia, muchos de los métodos y técnicas utilizadas en el transporte urbano no son aplicables o deben adaptarse al caso del transporte interurbano. Considerando estos aspectos y el alto impacto socio-económico que tiene el transporte público interurbano por carretera en Canarias, se decide enfocar esta tesis doctoral en este tipo de transporte público.

1.2. Objetivos

El principal objetivo que comparten los trabajos presentados en los artículos de esta tesis que se presenta por compendio, ha sido diseñar sistemas inteligentes de transporte basados en estándares, que faciliten la estimación de la demanda y que permitan evaluar la calidad del servicio que se presta a la ciudadanía por parte de una empresa de transporte público interurbano por carretera, utilizando exclusivamente datos de explotación y de planificación de la actividad. Como consecuencia de este requisito, los sistemas desarrollados no han requerido ningún despliegue de elementos específicos (hardware o software) para la obtención de datos, y por tanto los modelos y técnicas desarrolladas son susceptibles de ser utilizadas por empresas o autoridades reguladoras del transporte.

1.3. Metodología, técnicas y recursos utilizados

Mientras que en las últimas décadas muchos de los trabajos científicos se basaban en la simulación de fenómenos complejos, en la actualidad es habitual utilizar la exploración de datos [38]: se generan en multitud de sistemas y sensores, se depuran, almacenan, filtran y complementan, se procesan para la búsqueda de tendencias o patrones utilizando técnicas de Minería de Datos, y finalmente se analizan e interpretan, constituyendo lo que se denomina un proceso de descubrimiento de conocimiento [8].

Con el fin de garantizar la validez de los resultados obtenidos, todos los artículos presentados en este documento se han inspirado en el modelo de proceso no-propietario y documentado de Minería de Datos CRISP-DM [39], acrónimo de *Cross-Industry Standard Process for Data Mining*, que determina buenas prácticas para alcanzar mejores y más rápidos resultados en proyectos de esta naturaleza. Está constituido por las seis fases que de manera muy resumida se relacionan a continuación:

- **Comprensión del negocio:** Definir los objetivos a alcanzar en el proyecto, recopilando y analizando la información sobre los recursos disponibles y la prioridades de la organización participante.

- **Comprensión de los datos:** Estudiar de cerca los datos disponibles para minimizar los problemas que puedan surgir en la siguiente fase de preparación de los datos, identificando problemas de calidad y seleccionando aquellos interesantes para la formulación de las hipótesis de partida.

- **Preparación de los datos:** Construir el conjunto de datos a ser modelado: seleccionando, limpiando, fusionando y formateando los datos, y definiendo subconjuntos cuando sea necesario.
- **Modelado:** Encontrar patrones utilizando diferentes algoritmos y técnicas, y utilizando distintos métodos y parámetros hasta conseguir los mejores resultados, evaluando las soluciones obtenidas y volviendo a la fase de preparación de los datos si fuera necesario.
- **Evaluación:** Verificar que los resultados responden a las hipótesis planteadas inicialmente.
- **Implementación:** Aplicar las soluciones desarrolladas.

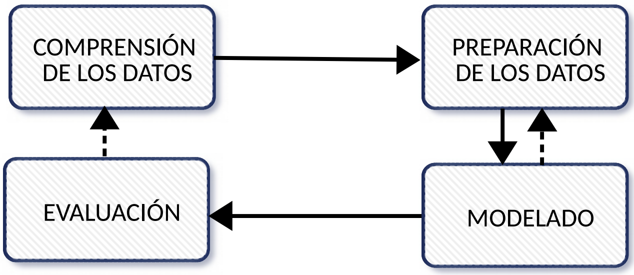


Figura 1: Fases consideradas del modelo CRISP-DM

La primera fase, de comprensión del negocio, se puede asimilar en este caso a los estudios y análisis previos a la ejecución de los trabajos de investigación contemplados en este documento; y no tiene cabida la última, la relacionada con la implementación en una organización. El resto de las fases que sí se han considerado, se presentan en la figura 1, donde puede observarse que el proceso no es estrictamente secuencial, las fases se encuentran relacionadas, se avanza (flechas de línea continua) o se retrocede (flechas punteadas) dependiendo de si los resultados alcanzan o no los objetivos planteados en cada una de ellas. Es necesario hacer notar, que alguna de las etapas de la metodología seguida en los artículos presentados, a su vez, puede estar constituida por un ciclo de fases como el mostrado, tal y como se podrá ver en detalle en la sección que contiene el resumen de los mismos.

De las fases detalladas, es indudable que la fase de Modelado es la que mayor importancia cobra en las referidas publicaciones, puesto que su carácter novedoso radica en la aplicación de distintas técnicas de Minería de Datos, concretamente en la aplicación de algoritmos de Inteligencia Artificial.

Pueden ser dos los objetivos en un proceso de descubrimiento de conocimiento con Minería de Datos [8]:

- describir los datos en forma de patrones inteligibles
- predecir el comportamiento futuro de algunas variables

y los dos están presentes en los trabajos publicados: buscando patrones de comportamiento de la demanda, identificando paradas, modelando los tiempos de paso de las expediciones y los tiempos de viaje, utilizando para ello técnicas de agrupamiento; y creando funciones para predecir la demanda de viajeros aplicando redes neuronales.

El agrupamiento es una técnica que realiza una distribución de elementos entre un número prefijado de grupos, particiones o segmentos, de acuerdo a una medida de similitud entre ellos, de tal forma que la similitud media entre elementos del mismo grupo sea alta y la similitud media entre elementos de distintos grupos sea baja. Un elemento que se suele utilizar como identificativo de un grupo es el centroide, que se define como aquel elemento (existente o no en el conjunto de datos inicial, dependerá de la técnica aplicada) que minimiza la suma de las similitudes al resto de los elementos del grupo. Para evaluar la calidad de los segmentos resultantes no es posible utilizar medidas basadas en la asignación previa de cada elemento a su grupo, como el índice de Jaccard. Una forma de hacerlo es en base a sus *siluetas* [40], que representan su "estrechez" y su "separación", utilizando el promedio de las siluetas de las particiones como medida para seleccionar el número óptimo de grupos en el conjunto de datos.

Las redes neuronales, en cambio, son modelos matemáticos muy genéricos, precisos, equivalentes al modelo autoregresivo no lineal para series temporales, y muy convenientes para abordar problemas de transporte por su capacidad de trabajar con cantidades masivas de datos multi-dimencionales, su flexibilidad de modelado y generalización [9]. Inspirado en el modelo biológico, se componen de una serie de elementos interconectados de procesamiento simple, denominados neuronas o nodos. Cada nodo recibe una señal de entrada que puede ser un estímulo externo o la señal de salida (información) de otros nodos, la procesa con una función de activación o transferencia, y finalmente genera una señal de salida, externa o hacia otros nodos.

Existe una gran variedad de redes neuronales, pero la más utilizada en problemas de estimación es la denominada de perceptrones multicapa (multi-layer perceptrones, MPL), compuesta de varias capas de nodos [41]:

- La primera capa, la de entrada, es la que recibe la información externa, las variables independientes o predictoras, las observaciones históricas de los datos.
- Las capas intermedias, o capas ocultas, constituidas por nodos completamente conectados.
- La última, la capa de salida, la que genera la solución al problema.

Una red neuronal debe ser entrenada para poder dar respuesta a un problema específico. Conocida la salida de un conjunto significativo de datos de entrada, el proceso de entrenamiento consiste en determinar el peso de los arcos de conexión entre nodos que minimice una determinada función de error en la salida, como por ejemplo, el error cuadrático medio. Una vez finalizado el proceso, el conocimiento aprendido se encuentra almacenado en el peso de los arcos y en el sesgo de los nodos que conforman la red.

Finalmente indicar que todos los desarrollos y herramientas utilizadas en los artículos presentados se han realizado con software de código abierto, y las principales aplicaciones han sido:

- Oracle SQL Developer [42], entorno de trabajo libre para el acceso y gestión de bases de datos.
- Pentaho Data Integration – Kettle [43], herramienta de la plataforma Hitachi Vantara que permite extraer y manipular datos de distintas fuentes para generar los eventos a tratar en las tareas de modelado.
- Rstudio [44]. Entorno de desarrollo de R [45], software que proporciona herramientas estadísticas y un considerable número de paquetes entre los que se incluyen gráficos, redes neuronales y algoritmos de agrupamiento.

1.4. Resumen de las publicaciones.

En esta sección se presenta un resumen de cada uno de los artículos que conforman esta tesis, en cuyas fases experimentales se utilizaron los datos de explotación de la actividad de transporte de la operadora SALCAI UTINSA S.A. (GLOBAL), empresa que tiene asignada la prestación del servicio público regular de viajeros interurbano por carretera por parte de la Autoridad Única del Transporte de Gran Canaria, siendo sus principales magnitudes [46]:

- ✓ 110 líneas
- ✓ 310 vehículos
- ✓ 2.400 expediciones diarias
- ✓ 2.700 paradas
- ✓ 25.000.000 km recorridos al año
- ✓ 20.000.000 de viajeros al año

1.4.1. Applying Time-Dependent Attributes to Represent Demand in Road Mass Transit Systems

Utilizando atributos dependientes del tiempo para representar la demanda en sistemas de viajeros por carretera.

El factor tiempo es determinante en cualquier análisis relacionado con el transporte de viajeros y son múltiples las maneras de tratarlo, tal y como puede apreciarse en los trabajos destacados en la sección Estado del arte. Cuando la atención se fija en la demanda y en la planificación de los servicios suele distinguirse entre días laborables, fines de semana y festivos, períodos lectivos y períodos de vacaciones, y dentro de un mismo día, horas picos y valles de afluencia de viajeros [15][25][26].

Atendiendo a la complejidad en el tratamiento de este factor tiempo se plantea la siguiente hipótesis:

Partiendo de los datos registrados en los viajes realizados por los usuarios del transporte público en un amplio período de tiempo, es posible caracterizar la demanda entre una parada origen y otra destino en un determinado intervalo temporal.

Con objeto de verificar la anterior hipótesis se plantea un nuevo tipo de datos, un nuevo atributo que haga posible:

- Simplificar los procedimientos de predicción de la demanda entre dos paradas, asociando ese valor a cada intervalo a estimar.
- Obtener información inteligible sobre el flujo de viajeros entre dos paradas que sirva de apoyo en las tareas de planificación de los servicios.

En este trabajo se propone y se justifica una metodología para la obtención del atributo propuesto, constituida por las siguientes fases:

1. Definición de datos. Identificación de las características temporales, de las escalas temporales a tratar en la generación del nuevo atributo.
2. Análisis y evaluación de las diferentes técnicas de inteligencia artificial para descubrir categorías, agrupamientos, en ese conjunto de datos, y definición de los procedimientos de evaluación de los distintos resultados de clasificación.
3. Definición y aplicación de los procedimientos de validación de los resultados obtenidos, y validación de la hipótesis de partida.

En la figura 2 se puede apreciar que cada una de las fases está compuesta a su vez por aquellas características de un proceso de descubrimiento de conocimiento, y se presenta además la interrelación existente entre ellas puesto que, como ya se ha comentado, no conseguir los resultados esperados supone retroceder y replantear las técnicas utilizadas y las decisiones tomadas.

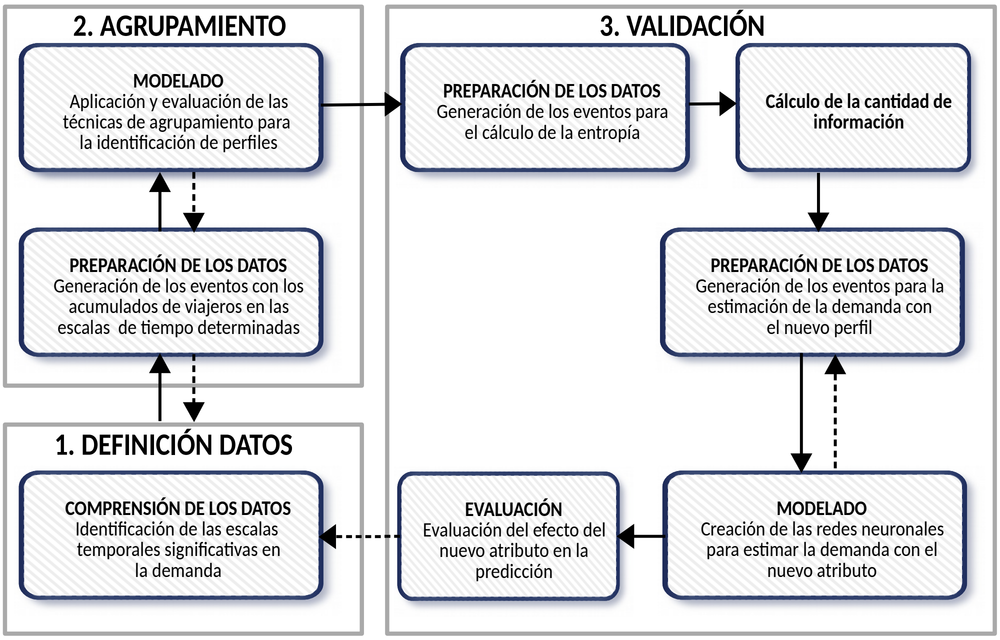

2. AGRUPAMIENTO
MODELADO
Aplicación y evaluación de las técnicas de agrupamiento para la identificación de perfiles
PREPARACIÓN DE LOS DATOS
Generación de los eventos con los acumulados de viajeros en las escalas de tiempo determinadas
1. DEFINICIÓN DATOS
COMPRENSIÓN DE LOS DATOS
Identificación de las escalas temporales significativas en la demanda
3. VALIDACIÓN
PREPARACIÓN DE LOS DATOS
Generación de los eventos para el cálculo de la entropía
Cálculo de la cantidad de información
PREPARACIÓN DE LOS DATOS
Generación de los eventos para la estimación de la demanda con el nuevo perfil
MODELADO
Creación de las redes neuronales para estimar la demanda con el nuevo atributo
EVALUACIÓN
Evaluación del efecto del nuevo atributo en la predicción

Figura 2: Fases realizadas en el artículo primero

Para la fase experimental se seleccionaron dos paradas (parada origen y parada destino) en tres de las principales rutas de la operadora de transporte, líneas con diferentes tipos de recorrido y de uso. Las conclusiones y resultados de cada etapa han sido:

1. Definición de los datos. Utilizando exclusivamente datos registrados por los sistemas instalados en los vehículos de la flota, concretamente el inicio y el final de la expedición y el pago realizado por el viajero, se estudian distintas combinaciones de valores temporales (semana del año, mes, día de la semana, día festivo, laborable, ...) que fueron posteriormente desechadas a la vista de los resultados obtenidos en las tareas de evaluación propias de las fases sucesivas. Finalmente se selecciona la siguiente escala de tiempo: la semana del año, el día de la semana y la hora del día.

2. Agrupamiento. Para cada escala de tiempo analizada, se generan los registros de datos con el total de viajeros que han ido de la parada origen a la parada destino en ese período de tiempo. La estructura final está constituida por los siguientes campos.

   - P_o parada origen de la demanda a analizar.
   - P_d parada destino de la demanda a analizar.
   - W número de semana del año.
   - D día de la semana.
   - $A_{W,D,H}$ demanda total en un día D, de la semana W, a la hora H desde la parada P_o a la parada $P_{d,}$ donde H puede tomar los valores $\{h_0, h_1, \ldots h_f\}$ siendo h_0 la primera hora del día analizada y h_f la última hora del día analizada.

   El conjunto de registros se somete a procedimientos de segmentación de 2 a *k* grupos que dan lugar a *k*-1 atributos, donde cada grupo aglutina aquellos registros con mayor similitud e identifican perfiles diferenciados de demanda. La validez de los resultados del procedimiento se evalúa midiendo la silueta de los conjuntos de registros generados para cada valor de *k* analizado.

3. Selección y validación de los segmentos generados en la fase anterior. En este paso se realizan dos evaluaciones adicionales: la primera aplicando un criterio independiente, calculando la ganancia de información de las distintas segmentaciones con los valores de k seleccionados en la fase anterior, y la

segunda aplicando un criterio dependiente, utilizando el nuevo atributo asociado al valor de k que aporta mayor información para estimar la demanda utilizando redes neuronales.

En todos los casos se consigue mejorar los resultados de predicción reduciendo el error observado, aunque el resultado ha sido desigual debido a las distintas características de las paradas tratadas, con flujos de viajeros más o menos estacionarios.

Por todo ello, se concluye que es posible generar un nuevo atributo temporal para clasificar la demanda en sistemas de transporte público por carretera, capaz de suministrar más información de los usados tradicionalmente, y representando nuevo conocimiento que ayude a entender el pasado y el presente, y prediga el futuro. La generación de este nuevo atributo se realiza usando como partida datos habitualmente utilizados por las empresas operadoras de transporte público, por lo que puede usarse en un análisis sistemático de la demanda, poniendo en evidencia aspectos tales como: paradas generadoras de demanda, demandas simétricas y no simétricas entre puntos de la red de transporte en función de la época del año o de la franja horaria del día, etc.

1.4.2. System Proposal for Mass Transit Service Quality Control Based on GPS Data

Propuesta de sistema para controlar la calidad del servicio de transporte público de viajeros basado en datos GPS

Que los vehículos lleguen a su hora a las paradas y a las estaciones es uno de los aspectos más relevantes a la hora de percibir la calidad del servicio por parte de los usuarios, pero el transporte urbano e interurbano por carretera se puede ver afectado por variables externas ajenas a la operadora, tales como densidad del tráfico, condiciones meteorológicas, cambios en la demanda de los viajeros, etc. Por todo esto, solo una evaluación continua de las funciones que miden la adherencia de los tiempos de paso por las paradas respecto a la planificación de las expediciones, ya sea por frecuencia o por horario de paso, hace posible un control exhaustivo de ese factor de calidad.

La principal aportación de este trabajo es la verificación de una hipótesis que consiste en que, partiendo únicamente de los datos de posicionamiento que habitualmente se registran en los los vehículos de la flota de transporte público, es posible realizar un control exhaustivo de la calidad en el cumplimiento horario a partir de una evaluación continua de las métricas de calidad correspondientes. El resultado de dicha evaluación puede ser utilizado para obtener nueva información valiosa de la red de transporte. La descripción de la metodología y técnicas utilizadas para la verificación de esta hipótesis es otra aportación de este trabajo. Al igual que en el trabajo anterior, dado que los datos utilizados son datos habitualmente manejados por las empresas operadoras de transporte público, la metodología propuesta puede ser utilizada sin requerir ningún despliegue adicional de elementos hardware o software en los vehículos.

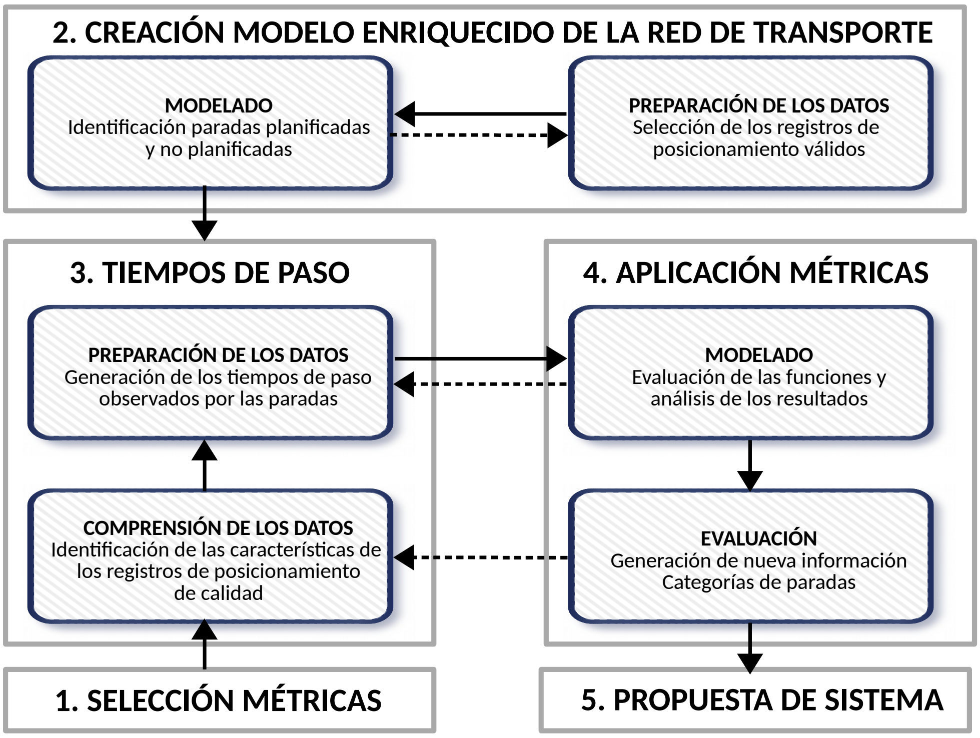

2. CREACIÓN MODELO ENRIQUECIDO DE LA RED DE TRANSPORTE
MODELADO
Identificación paradas planificadas y no planificadas
PREPARACIÓN DE LOS DATOS
Selección de los registros de posicionamiento válidos
3. TIEMPOS DE PASO
PREPARACIÓN DE LOS DATOS
Generación de los tiempos de paso observados por las paradas
COMPRENSIÓN DE LOS DATOS
Identificación de las características de los registros de posicionamiento de calidad
4. APLICACIÓN MÉTRICAS
MODELADO
Evaluación de las funciones y análisis de los resultados
EVALUACIÓN
Generación de nueva información Categorías de paradas
1. SELECCIÓN MÉTRICAS
5. PROPUESTA DE SISTEMA

Figura 3: Fases realizadas en el artículo segundo

Para alcanzar este objetivo se han ejecutado las siguientes fases cuyas relaciones se muestran en la figura 3:

1. Análisis y selección de las funciones, de las métricas, relacionadas con la calidad del tiempo de paso en transporte por carretera.

2. Creación del “Modelo Enriquecido de la Red de Transporte”, necesario para disponer de la ubicación de las paradas de las líneas a analizar [47]

3. Análisis y selección de las características del “Conjunto de Registros de Posicionamiento de Calidad”, que sirva de base para obtener los tiempos de paso de las expediciones realizadas en una línea en un período de tiempo determinado. Las expediciones cuyo recorrido no pueda reconstruirse con registros de esas características serán ignoradas.

4. Aplicación de las métricas de calidad seleccionadas al conjunto de datos, análisis de los resultados y verificación de la obtención de nueva información.

5. Propuesta de sistema de control de calidad.

Aplicando las tareas propuestas a una línea concreta de la red de transporte, a continuación se presentan las principales decisiones tomadas y los resultados alcanzados en cada una de ellas.

1. Debido a la naturaleza de los datos disponibles, transporte planificado de baja frecuencia, se seleccionaron dos de los indicadores comúnmente más utilizados para medir la confiabiliad en planificación [11]:

- “Retraso en la Llegada” (*Arrival Delay* – AD) que representa, de cada parada de cada expedición realizada, la diferencia entre el tiempo de llegada real y el tiempo de llegada planificado.

$$AD_{\text{parada i}} = \text{tiempo de paso real}_{\text{parada i}} - \text{tiempo de paso planificado}_{\text{parada i}} \tag{1}$$

- “Variación de Tiempo de Viaje” (*Run Time Variati*on – RVT), de especial interés para trayectos de larga distancia, calcula la variación de tiempo que se produce en cada expedición. Depende del número de paradas total y del retraso en la llegada a cada parada.

$$RTV_{\text{expedición e}} = \frac{\sum_{i=1}^{n} \frac{\left|\text{tiempo de paso real}_{\text{parada i}} - \text{tiempo de paso planificado}_{\text{parada i}}\right|}{\text{tiempo de paso real}_{\text{parada i}}}}{n} \tag{2}$$

2. Puesto que se parte únicamente de los registros de posicionamiento registrados en los vehículos, es necesario obtener en primer lugar la ubicación exacta de las paradas, indispensables para disponer del tiempo de paso real por cada una de ellas. Es por eso por lo que, partiendo del conjunto

de posicionamientos de calidad registrados con velocidad cero, con el vehículo parado, y aplicando técnicas de agrupamiento, se genera el Modelo Enriquecido de la Red de Transporte que distingue entre aquellas paradas planificadas en el servicio y aquellas relacionadas con señales de tráfico. Para el análisis realizado en este artículo la ubicación de las paradas está determinado por los centroides de cada uno de los grupos asociados a paradas planificadas.

3. Un aspecto fundamental para un control de calidad de confianza es la integridad de los datos que participan en el análisis, por lo que para obtener los tiempos de paso de cada expedición por las paradas que conforman la ruta, el conjunto de registros de posicionamiento a considerar de cada una de ellas ha de cumplir las siguientes condiciones:

    - Ha de contener datos de posicionamiento de buena calidad.
    - Debe ser coherente con las expediciones realizadas:
        - Representar todo trayecto realizado, de inicio a fin.
        - Contener al menos un registro de posicionamiento para cada parada de la ruta cuya distancia a la ubicación determinada por el centroide correspondiente sea menor o igual a un umbral previamente definido, y que determina la precisión de las medidas.

    Seleccionadas las expediciones que cuentan con registros de posicionamiento de calidad, se calculan los tiempos de paso por cada una de las paradas de la ruta, que viene determinado por el momento en el que se registra la localización más cercana a su ubicación real.

4. Una vez que se dispone de las horas de paso por las paradas de un conjunto significativo de expediciones para el período de tiempo de estudio, se evalúa la puntualidad o la regularidad de los servicios aplicando las métricas de calidad. Cuando la función es el “Retraso en la Llegada”, el resultado dará lugar a nuevas categorías de paradas, que supone una evaluación de la calidad en el espacio. Cuando se analiza la “Variación de Tiempo de Viaje” de las expediciones, se realiza una evaluación de la calidad en el tiempo.

5. De manera resumida, la propuesta final de sistema de control de la calidad se muestra en la figura 4, donde se destaca especialmente que la única fuente de datos que se considera para el control de la calidad son los registros de

posicionamiento generados en los vehículos y almacenados en la base de datos de transporte de la operadora.

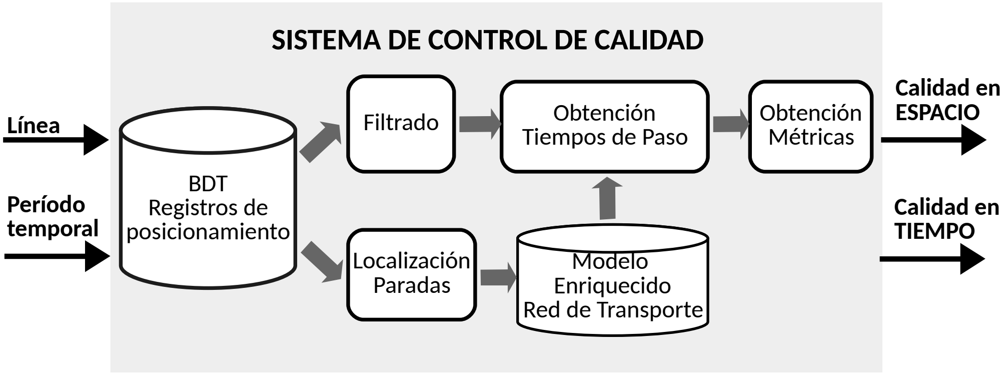

SISTEMA DE CONTROL DE CALIDAD
Línea
Período temporal
BDT Registros de posicionamiento
Filtrado
Obtención Tiempos de Paso
Obtención Métricas
Localización Paradas
Modelo Enriquecido Red de Transporte
Calidad en el ESPACIO
Calidad en el TIEMPO

Figura 4: Sistema de calidad propuesto

Teniendo en cuenta los resultados obtenidos, se concluye que es posible diseñar un sistema de control de calidad mediante la evaluación continua de los tiempos de paso por las paradas de la red de transporte. Un sistema que además detecta aquellos lugares de la ruta que afectan especialmente a la duración y variabilidad del tiempo de viaje, uno de los principales indicadores de calidad de servicio en el transporte público de viajeros. La identificación de estos puntos juega un papel importante en el paradigma actual de gestión inteligente del tráfico, ya que permite la aplicación de actuaciones específicas orientadas a agilizar el tránsito de los vehículos de transporte público como puede ser, por ejemplo, la instalación de sistemas de prioridad semafórica.

1.4.3. Systematic Approach to Analyze Travel Time in Road-based Mass Transit Systems Based on Data Mining

Enfoque sistemático de análisis del tiempo de viaje del transporte público por carretera basado en Minería de Datos

Como se ha comentado con anterioridad, para que el transporte público sea una alternativa factible a los vehículos privados, éste debe ofrecer servicios de calidad, atractivos para el ciudadano. Especialmente cuando se trata de transporte interurbano el tiempo de viaje es un factor de calidad relevante, desde el punto de vista de la planificación y el control, y como parte fundamental de la información que se ha de suministrar a los viajeros. Por todo ello, para las operadoras del transporte es importante poder disponer de herramientas que les permitan evaluar y analizar los tiempos empleados en los viajes con objeto de ajustar las planificaciones, analizar y corregir los factores que afecta a su variabilidad, y mejorar la información dirigida a los usuarios.

La principal aportación de este artículo es la verificación de la siguiente hipótesis: a partir de la observación de lo que sucede en la red de transporte, partiendo de los registros de posicionamiento registrados por los vehículos junto con los datos de planificación de una línea, es posible realizar un análisis continuo de los tiempos de viaje invertidos por las expediciones con objeto de identificar su comportamiento y localizar los factores que afectan a su variabilidad. Al igual que en las dos publicaciones anteriores, otra aportación es la descripción de la metodología y técnicas utilizadas para la verificación de esta hipótesis.

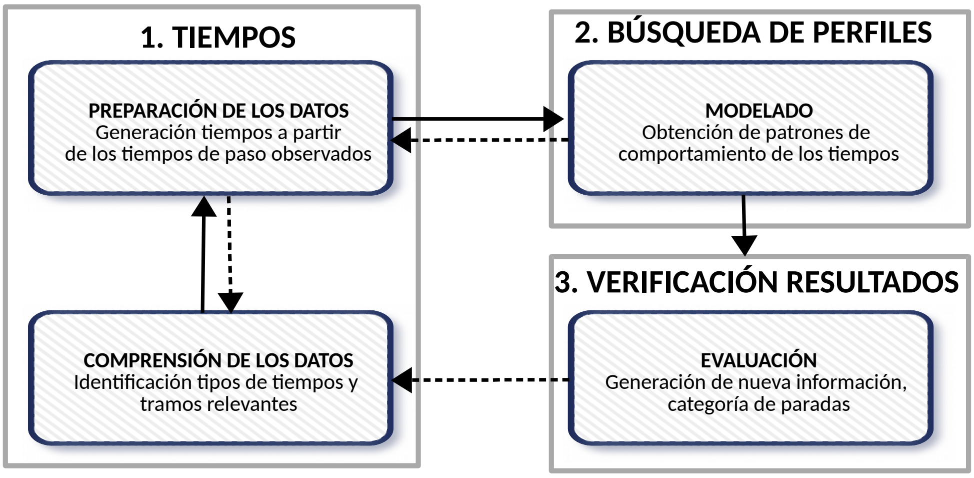

1. TIEMPOS
PREPARACIÓN DE LOS DATOS
Generación tiempos a partir de los tiempos de paso observados
COMPRENSIÓN DE LOS DATOS
Identificación tipos de tiempos y tramos relevantes
2. BÚSQUEDA DE PERFILES
MODELADO
Obtención de patrones de comportamiento de los tiempos
3. VERIFICACIÓN RESULTADOS
EVALUACIÓN
Generación de nueva información, categoría de paradas

Figura 5: Fases realizadas en artículo tercero

De manera esquemática se plantean las siguientes fases, que se presentan en la figura 5:

1. Definición de los tiempos y los tramos a analizar. Definición y obtención de los tiempos de paso a partir del "Conjunto de Registros de Posicionamiento de Calidad" y generación de los conjuntos de datos a modelar.
2. Determinación de las técnicas de extracción de patrones y de los métodos de evaluación de resultados.
3. Determinación de procedimientos de evaluación de la nueva información.

Aplicando la metodología propuesta a la misma línea analizada en el artículo presentado en el apartado anterior, se tomaron las siguientes decisiones y se alcanzaron los siguientes resultados:

1. En esta primera fase se abordan dos cuestiones primordiales: la primera relacionada con la definición de los tiempos de interés para el objetivo planteado, y la segunda relacionada con la ruta a analizar:

   I. Tiempos. Siendo fundamentalmente dos los objetivos a alcanzar, conocer la variabilidad de los tiempos de viaje, su comportamiento en distintos períodos temporales, y descubrir los factores que le afectan, se decide analizar los siguientes:

   - Tiempo Observado de paso de cada expedición por cada una de las paradas a analizar.
   - "Retraso en la Llegada" (AD) de cada expedición por cada una de las paradas, función utilizada habitualmente en el ámbito del transporte que se presentó en el artículo anterior, ecuación (1).
   - "Retraso Relativo en la Llegada" (*Relative Arrival Delay* – RAD) respecto a la parada anterior. Con esta medida de tiempo se pretende evaluar lo que tardan los vehículos entre dos paradas consecutivas de tal forma que, si el valor es positivo significa que han empleado más tiempo que el planificado, y si es negativo que ha empleado menos tiempo que el planificado. De esta manera los tramos más conflictivos destacan de manera inmediata. Este tiempo viene dado por la siguiente fórmula:

   $$RAD_{\text{parada i}} = AD_{\text{parada i}} - AD_{\text{parada i-1}} \quad (3)$$

   II. Tramos. Las líneas en el transporte público interurbano cubren trayectos de muchos kilómetros y tienen definidas muchas paradas. Además, algunos de estos trayectos y paradas se encuentran dentro de zonas urbanas, otros en vías rápidas, y otros en vías comarcales. Por otro lado, hay paradas con gran afluencia de viajeros, otras donde la proporción respecto al total de la línea es poco significativa e incluso alguna que apenas tiene usuarios a lo largo de la semana. Por todo esto, y con el ánimo de que la información suministrada por el sistema fuera lo más inteligible posible, se decide no considerar todas las paradas definidas en la línea sino seleccionar aquellas que cumplan alguno de los siguientes criterios:

- tengan un tránsito de viajeros de al menos un 10% del total de usuarios de la línea, o
- se encuentren ubicadas al comienzo o al final de un tramo donde se produce un cambio de vía

Una vez definidas estas cuestiones, y partiendo de registros de posicionamientos de la expediciones completos y coherentes, con las características que ya fueron detallas en el artículo anterior, se genera un conjunto de datos para cada una de las medidas de tiempo consideradas de interés: Tiempo Observado, Retraso en la Llegada y Retraso Relativo.

2. El objetivo de esta fase es encontrar patrones de comportamiento de las expediciones en lo que se refiere a los tiempos seleccionados, aplicando para ello técnicas de segmentación y de evaluación. De los resultados obtenidos no solo se concluye el número óptimo de agrupaciones a analizar en la etapa siguiente, sino también la conveniencia de utilizar determinada medida de tiempo o determinados tramos de la ruta, atendiendo a la información que aportan los centroides.
3. En esta etapa de evaluación final se utilizan tablas de contingencia para analizar la relación entre las expediciones incluidas en cada segmento de cada una de las medidas de tiempo consideradas, y distintos períodos temporales (hora del día, día de la semana, mes). Dada la naturaleza de dichas medidas, se puede obtener información relativa a cuándo las expediciones sufren retrasos respecto al horario planificado, y también información respecto a dónde se producen esos retrasos y qué tramos de la ruta afectan al incumplimiento horario.

Por lo tanto, atendiendo a los resultados obtenidos, puede concluirse que es posible localizar los factores que afectan a la variabilidad de los tiempos de viaje en el espacio y en el tiempo partiendo de los registros de posicionamiento y de la planificación de las líneas. Además, la metodología utilizada basada en Minería de Datos, se puede aplicar de manera sistemática con el fin de garantizar un nivel de cumplimiento de los horarios de paso aceptable para el usuario de transporte público, y de facilitar la predicción del tiempo de viaje en función de aspectos tales como época del año, días de la semana y franja horaria del día. Esta metodología introduce como aspecto novedoso el uso de las técnicas de agrupamiento con el fin de obtener los patrones de comportamiento del tiempo de viaje.

1.5. Coherencia y unidad temática.

Todos los artículos presentados en este documento han sido realizados dentro de la línea de investigación sobre “Sistemas de Transporte Inteligentes” de la división “Sistemas de Información Móviles” del Instituto Universitario de Ciencias y Tecnologías Cibernéticas de la Universidad de Las Palmas de Gran Canaria, en colaboración con la empresa SALCAI UTINSA S.A. (GLOBAL), que en los últimos años ha puesto a disposición del equipo investigador datos de explotación de la actividad del transporte que realiza. Fruto de esa estrecha colaboración se ha de destacar el trabajo publicado que define un marco de desarrollo sistemático de STI para el transporte público de carretera [48].

Como se ha comentado con anterioridad, han sido dos los objetivos perseguidos en los trabajos que constituyen esta tesis - sin perder de vista los intereses expresados por la empresa - por un lado, la predicción de la demanda, y por otro la evaluación de la calidad del servicio. Ambos han sido abordados desde la perspectiva de la incorporación de técnicas de Inteligencia Artificial como procedimientos dentro de un proyecto de Minería de Datos, y tratados en paralelo:

- Análisis de la demanda. En este caso, el propósito es encontrar un patrón para predecir el número de pasajeros que querrán ir de un lugar a otro en un momento determinado sin considerar la planificación de lineas y horarios. En primera instancia y en una publicación anterior se consigue modelar el comportamiento en un corredor determinado utilizando algoritmos de clasificación [49], y es en ese momento cuando, analizando las publicaciones existentes en ese ámbito, se plantea el interés de encontrar una nueva manera de representar el tiempo que facilite la predicción de la demanda, y es la aplicación de técnicas de agrupamiento y de redes neuronales a ese problema, lo que da origen al primer artículo que se presenta en este documento.

- Calidad del servicio. En base a los datos registrados por los sistemas de posicionamiento, el segundo de los artículos es el resultado de utilizar técnicas estadísticas y de reconocimiento de patrones [47] para proponer un sistema de control de la calidad de los servicios de transporte. El tercero de los artículos, ha de considerarse como consecuencia de los dos primeros ya que tiene como objetivo medir la calidad del servicio analizando los tiempos de viaje con las técnicas de agrupamiento utilizadas previamente en el análisis de la demanda.

2. Publicaciones originales

2.1. Applying Time-Dependent Attributes to Represent Demand in Road Mass Transit Systems

entropy

Article

Applying Time-Dependent Attributes to Represent Demand in Road Mass Transit Systems

Teresa Cristóbal [1], Gabino Padrón [1], Javier Lorenzo-Navarro [2], Alexis Quesada-Arencibia [1] and Carmelo R. García [1,*]

[1] Institute for Cybernetics, Campus de Tafira, Las Palmas de Gran Canaria, University of Las Palmas de Gran Canaria, 35017 Las Palmas, Spain; teresa.cristobal@fpct.ulpgc.es (T.C.); gabino.padron@ulpgc.es (G.P.); Alexis.quesada@ulpgc.es (A.Q.-A.)

[2] University Institute of Intelligent Systems and Numeric Applications in Engineering, Campus de Tafira, Las Palmas de Gran Canaria, University of Las Palmas de Gran Canaria, 35017 Las Palmas, Spain; javier.lorenzo@ulpgc.es

* Correspondence: ruben.garcia@ulpgc.es; Tel.: +34-928-458-651; Fax: +34-928-458-700

Received: 24 January 2018; Accepted: 16 February 2018; Published: 20 February 2018

Abstract: The development of efficient mass transit systems that provide quality of service is a major challenge for modern societies. To meet this challenge, it is essential to understand user demand. This article proposes using new time-dependent attributes to represent demand, attributes that differ from those that have traditionally been used in the design and planning of this type of transit system. Data mining was used to obtain these new attributes; they were created using clustering techniques, and their quality evaluated with the Shannon entropy function and with neural networks. The methodology was implemented on an intercity public transport company and the results demonstrate that the attributes obtained offer a more precise understanding of demand and enable predictions to be made with acceptable precision.

Keywords: clustering; entropy; attribute creation; data mining; intelligent transport systems; mass transit systems; demand

1. Introduction

According to the International Energy Agency, there were an estimated 870 million passenger light-duty vehicles on our roads worldwide in 2011, a figure that is projected to grow to 1.7 million in 2035 [1]. This type of mobility, based on private vehicles, is resulting in the deterioration of health, the environment and safety on the roads. To illustrate this problem, the World Health Organisation estimates that approximately 3 million people die every year due to health problems caused by pollution [2] and in the European Union 250,000 people are victims of traffic accidents, and about 10% of these accidents are fatal [3]. Large-scale public road transport systems are an effective means to respond to mobility needs in a way that is safer and more respectful towards our health and the environment. For this reason, the development of efficient transportation systems that provide quality of service is a priority for the authorities and for transport agencies. Consequently, models and techniques that contribute to the development of efficient systems and that provide quality of service are a topic of great interest for the academic community.

In the context of large-scale road transport systems, knowledge of demand is important, as it must be taken into account when rolling out transport systems that provide quality of service to users and that are efficient from the point of view of resource requirements. Because demand in this type of system is variable, planning will vary depending on time-dependent characteristics, such as: time of year, month and day, etc. This article proposes a new type of time-dependent attribute that will enable a more precise understanding of demand, and describes how to obtain the attributes systematically by means of data mining.

This paper makes three main contributions. The first is the proposal of a new class of attributes to represent the behaviour of demand. These attributes are sensitive to the time periods that affect demand (time of year, month, day of the week, work day, public holiday, time of the day, etc.). They provide information about demand and therefore help to develop schedule plans adapted to the behaviour and to predict future demand. The second contribution is the fact that the proposed attributes are the result of a clustering process, in which each cluster represents a different demand pattern occurring in different instants of time. This is a novel approach to classification. The third contribution is the fact that it was developed in a public road transport system planned by timetable, an approach used in intercity or long-distance transport. Most of the studies that have addressed problems related to the subject matter of this paper have been developed in the context of systems planned by frequency of stops, which is the case of urban transport systems.

The rest of this article is organised into five sections. The following section will review studies related to the proposed methodology, which will serve to contextualise and highlight its interest and originality. The methodology, based on data mining, used to obtain the new representation model is described in the third section. The fourth section is devoted to presenting the results obtained in a real use case, as the representation model has already been implemented in a public passenger transport company. The results are discussed in the fifth section. Finally, the conclusions are presented in the sixth section.

2. Related Studies

The review in this section aims to meet two objectives. The first is to highlight the importance of demand and models that represent it when developing mass transit systems that provide a quality service. The second is to review the use of data mining in public transport, setting out the goals that were pursued and the techniques used.

The efficiency and quality of service in mass transit systems is an important challenge for the authorities and for transport agencies. This challenge should be addressed through the use of methodologies and techniques that allow optimal design of the transport network, of planning and operations monitoring. Moreira [4] conducted an exhaustive review of these techniques and methodologies, often in relation to demand.

The solution to the problem of designing an optimal transport network resides in striking a balance between normally opposed factors. A first, essential factor is responding to the mobility needs of users. Another factor, related to the viability and sustainability of the network, is the resources that are used. A final factor is the time spent travelling: travel time. In the various methodologies proposed for an optimal design of the transport network, demand is a factor to consider, and the most widely-used form of representation is the origin-destination matrix (O–D matrix). This representation method consists of a matrix in which each row represents the trip origin nodes and the columns the destination nodes. Thus, the value $M(x,y)$ represents the number of travellers who have travelled from node x to node y in a given period of time (Desaulniers [5]). In this context, different demand models have been developed, which can be classified into two categories: deterministic and stochastic. Deterministic models have been used to predict demand over a period of time [6]. Stochastic models are a variant of deterministic models. They are used to predict the demand variations that occur daily in the transport network, for example, and are represented by the O–D matrix containing the mean value and the standard deviation [7]. An alternative approach to determining passenger flow using the O–D matrix was proposed by Berbey [8], using fuzzy logic to determine the values of its elements on a public transport metro system.

In the context of mass transit systems by road, the purpose of an operations planning design problem is to specify a sequence of operations to be carried out by vehicles in order to create a quality transport service. In this context, quality of service means adherence to the planned schedules and reduction of traveller waiting times at stops. There are two types of operations planning: planning by frequency of stops and planning by timetable. Planning by frequency of stops is used in urban transport and the models used generally do not take into account demand quantification data; they assume that the traveller goes to stops randomly. An exception to this approach is the work of Patnaik [9], who used data mining, specifically clustering techniques and decision trees, to develop optimal operations plans, using the cumulative number of passengers travelling in a vehicle on a trip and the passing times at each stop on the route. Using machine learning techniques—specifically, clustering techniques and decision rules—Mendes-Moreira [10] described a framework that uses automatic vehicle location systems (AVL) data to test whether the established schedule plan fits the conditions in the design of the transport network planned by frequency. Khiari [11] also used clustering techniques based on Gaussian Mixture Models (GMM) to set bus schedule coverage in a context of planning by frequency, using AVL and APC data. Although these two studies made use of clustering techniques, the techniques that they employed were different, as were the data used (location and passenger counting). When the route is planned by timetable, it is assumed that the arrival of passengers at the stop is a function of the scheduled passing time, namely, demand conditioned by service (Furth [12]); it is therefore necessary to predict the number of passengers that will be at the stops on a route at any given time.

On public road transport systems, operations monitoring aims to provide the appropriate response, usually in real time, to incidents that occur on the transport network and this may affect adherence to scheduled operations and quality of service. According to Dessouky [13], the type of information required by the methods implemented by these strategies are: planning of vehicle operations, information on vehicle headways, a prediction of headway times and a prediction of the passengers waiting at each of the stops on the route. For this type of method, it is assumed that, in general, the behaviour of the transport network is stochastic, using probabilistic models to represent different relevant factors involved in operations control. Hadas [14] proposed an optimisation method with the goal of improving the reliability of public transport services by optimally reducing the transfer times required in transport network operations, using O–D matrices obtained at a given time as input data in simulations to verify the validity of the proposal. Sáez [15] proposed a predictive control method, the objective of which was to minimise the passenger's travel time, modelling the passengers that boarded and alighted from the buses with O–D matrices obtained from historical data in different time periods.

Technological advances in public road transport systems, especially in on-board systems, sensor networks and personal devices, have allowed for a large amount of data. These data—from payment systems, automatic passenger counting systems (APC) and automatic vehicle location systems (AVL)—have allowed researchers to use data mining techniques and, especially, classification techniques to acquire information on the behaviour of the traveller and the transport network. In the case of the data from payment systems, particularly those provided by systems based on contactless cards, Agar [16] defined market segments within public transport users. Lathia [17] proposed a system to predict individual travel times and Du [18] also used the data from automatic payment systems, together with socio-demographic data, to obtain behaviour mobility patterns that would enable the demand for public transport services in areas of urban expansion to be estimated.

Using AVL systems as a data source, Sun [19] proposed a system to predict the arrival time of buses at stops by combining clustering techniques with other methods. Based on positioning data and, more specifically, on data generated by GPS devices on taxis, some studies have generated mobility patterns in urban areas. Yuan [20], starting from segmentation of such a space based on its main routes, and considering the points of interest located in each of those segments (restaurants, shopping centres, residential buildings, etc.) and the mobility patterns obtained for all the trips, obtained a set of functions for each of these segments that characterise this mobility. To this end, a model inspired by probabilistic topic models [21] was used to recognise content in documents. Zhao [22] segmented the urban space by means of the quad-tree division, considering only the population mobility data generated in taxis, without prior information on infrastructures, since a strong correlation had been found between infrastructure and the trips made by this means of transport.

3. Methodology

This paper was developed in the context of public intercity or long-distance road transport systems, which, as previously stated, are scheduled by timetable. In this type of system, operations planning is carried out using attributes that represent periods of time that must be adapted to variations in demand in order to optimise the resources to be used. Examples of commonly used time periods are: working days, public holidays, weekends, school day, etc. Therefore, information about the different periods that affect demand on the routes of the transport network is a basic requirement for optimal planning. For example, demand on one route will also vary on the same day, depending on the time of day. For this reason, scheduling should also use attributes that represent period types such as, for example, peak and off-peak. Therefore, to design an optimal schedule and to monitor that schedule, it is necessary to know how demand varies on a given day, and this demand depends on the time of day.

This paper proposes the use of a new type of attribute to classify demand. For a given route, each attribute value represents a demand pattern that may occur in different periods of time, such as: time of year, month, week, day of the week, time of day, etc. The aim is to obtain more precise knowledge of how demand varies over time, so that this knowledge can be used to optimise the planning and monitoring of operations. This section presents the methodology used to obtain these attributes. Based on data mining, it uses clustering techniques to generate them, entropy functions to evaluate the amount of information that they provide and neural networks to analyse their capacity to predict demand.

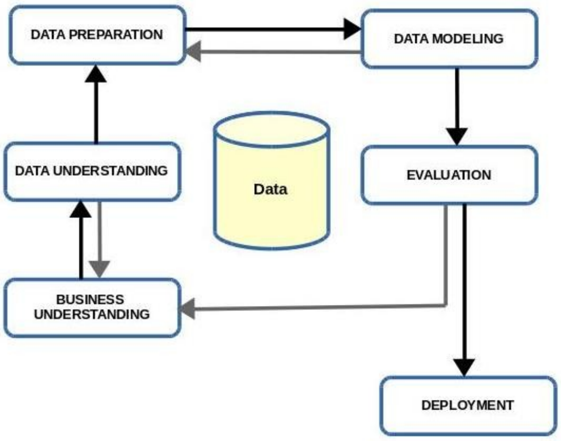


Figure 1. Process cycle of the CRoss-Industry Standard Process for Data Mining (CRISP-DM) model.

The methodology used was the process-oriented methodology called CRoss-Industry Standard Process for Data Mining (CRISP-DM) [23]. With this methodology, the processes are clustered in different phases to form the main cycle (see Figure 1). As can be seen, the process is not strictly sequential: the phases are interrelated and may move forward (black lines in the diagram) or backward (grey lines) depending on whether or not the objectives are attained. The purpose of the initial phase, Business Understanding, is to determine the scope of the problem and set the main data mining project targets. The data understanding phase is designed to analyse the available data and solve the problems that may be detected. The data preparation phase comprises all the tasks related to the construction of the final dataset that will be subjected to the procedures of the next phase, Modelling, in which different methods and parameters are tested, the validity of the results is evaluated and, if necessary, the data preparation phase begins anew. The evaluation phase verifies that the results respond to the needs of the organisation, which were defined in the first phase of the project, reconsidering them if necessary. Depending on the project requirements, tasks related to the final phase, Deployment, may differ, depending on the nature of the project, and involves incorporating the acquired information in the form of a report or as a new procedure in the organisation.

This study mainly encompasses two of these phases: Data preparation and Modelling. In the first, Data Preparation, the data that form the basis for this study (listed below) were merged and complemented, and the dataset for training and testing the neural networks was defined. In the second, Modelling, the modelling tools were incorporated, and different methods and parameters were tested to achieve the desired results.

The implementation of this methodology is described below. The public transport company Global Salcai-Utinsa collaborated with this study by providing access to their data. It is a company that operates on the island of Gran Canaria (Canary Islands, Spain) and is the main intercity transport company on this island; it has a fleet of 304 vehicles operating on a transport network with 2686 stops, 110 different routes and 2395 daily routes. Annually, its vehicles travel 28,897,002 kilometres and transport 19,284,378 passengers [24].

3.1. Data Preparation Phase

The input data used for this methodology are intrinsic to transport services and do not require external sources. Only records that are considered in the operational systems (operations and payments made on vehicles) are used, and others, such as vehicle location, are not included. Therefore, the method may be reproduced on most public transport companies since it does not require very sophisticated technological means. The data used came from the systems installed on the vehicles of the fleet. These systems record all the relevant events that occur on the vehicles. For this methodology, the data records are used to record the beginning and end of each line service, the change of stop, and the payment made by a traveller on the vehicle, which can be in cash or by means of a contactless card. Based on the records associated with these events, the objective is to obtain new attributes that describe the different passenger demand patterns and thus be able to estimate them. Considering the different periods of time that are used in the planning and monitoring of operations, to classify the demand patterns, it was considered sufficient to define three different time scales:

- Week of the year, to detect variations in demand depending on the time of year.
- Day of the week, to detect variations in demand depending on the day of the week.
- Time of day, to detect variations in demand depending on the time of day.

The events and data used in this study were as follows:

- Line service start event: the service start date and time, the vehicle that performed the line service, the route and the trip number.
- Line service end event: the line service end date and time, the vehicle that performed the line service, the route and the trip number.
- Stop change event: the date and time of the stop change, the vehicle that recorded the change of stop, the stop, route and trip number.
- Cash payment for travel event: the date and time of payment, the vehicle, the type of fare applied, the number of passengers, the origin stop of the trip, the destination stop of the trip, the cost, the route and the trip number.
- Contactless card payment for travel event: the date and time of payment, the vehicle, the type of fare applied, the number of passengers, the origin stop of the trip, the destination stop of the trip, the cost, the route and the trip number.

It should be noted that, although the events and data directly related to the purpose of this study are those that describe each trip (origin, destination, date, time and number of travellers), all other events and data have been used for validation purposes. It should also be noted that these data are operational data that are usually recorded by public transport companies; therefore, the proposed methodology may be applied to most intercity or long-distance transport companies without the need for any specific technological installation.

From these data, the initial dataset of records was constructed, specifying the point of origin—stop P_o on the transport network—the destination—stop P_d on the network—the demand and the period of time, T, to be analysed. With these initial specifications and accessing the transport operator's database, all the records that represent the line services from the period T and that passed through points P_o and P_d were obtained.

Once the integrity of the initial dataset was guaranteed, it was processed and two datasets were obtained. The first one was used to obtain the new attributes. The second was used as the training and test dataset of the demand prediction neural network. In Table 1, the record structure of the first dataset is represented; the meaning of the fields is as follows:

- P_o origin stop of the demand to be analysed.
- P_d destination stop of the demand to be analysed.
- W number of the week of the year, assuming that one year has 52 or 53 weeks.
- D day of the week: 1 Monday, 2 Tuesday, ..., 7 Sunday.
- $A_{W,D,H}$ total demand on day D, of week W, at time H from stop P_o to stop P_d. H_0 indicates the first hour of the day analysed and H_f indicates the last hour of the day analysed.

Table 1. Structure of the dataset for classifying demand.

P_o	P_d	W	D	$A_{W,D,H0}$	$A_{W,D,H1}$	$A_{W,D,H2}$	....	$A_{W,D,Hf}$

The steps followed to obtain the dataset for classifying demand are described below:

1. Data were obtained from the transport database relating to the bus lines operating on the routes that include the trip to be analysed in the time period T. This dataset was called L.
2. For each bus line in dataset L, data on the line services that operated during the time period T were obtained. This dataset was called SL.
3. For each line service in SL, data for all the payments made by the travellers for whom the origin stop was P_o and the destination stop P_d were obtained. This dataset was called MT.

For each payment record from the MT dataset, using the recorded date and time of payment, the movement was added to the corresponding records of the dataset for classifying demand.

The record structure of the training and test datasets for evaluation of the new attribute, based on its use in the prediction of demand using a neural network, is presented in Table 2. The meaning of the fields is as follows:

- P_o origin stop of the demand analysed.
- P_d destination stop of the demand analysed.
- W number of the week of the year, assuming that one year has 52 or 53 weeks.
- D day of the week: 1 Monday, 2 Tuesday, ..., 7 Sunday.
- F_{Po} indicates public holiday at the origin stop.
- F_{Pd} indicates public holiday at the destination stop.
- C is the new attribute, the value of which corresponds to a demand pattern for the values P_o, P_d, W and D, obtained in the clustering process applied to the dataset for classifying demand. The demand pattern is the identifier of the cluster.

- N is the number of passengers to be forecast, who go from stop P_o to stop P_d.

Table 2. Structure of the data records for the training and test dataset.

P_o	P_d	W	D	F_0	F_d	H	C	N

3.2. Modelling Phase

This phase consisted of two processes. The objective of the first was to obtain the new attributes that describe demand and the purpose of the second was to evaluate the attributes that were generated.

The task of obtaining the attributes was carried out using clustering techniques in order to group the data according to their similarity, thus enabling the creation of different categories. Specifically, for each pair of stops, a subset of data was generated that exclusively contained the demand data represented in Table 1 ($A_{W,D,H0}$, $A_{W,D,H1}$, ..., $A_{W,D,Hf}$). Each of the mentioned subsets was segmented into 2 to k clusters giving rise to new k-1 attributes, so that each value of the attribute corresponds to each of the clusters, which represent different demand patterns. There are many segmentation algorithms that Xu [25] classified depending on methodology and philosophy: those that use measures of distance and similarity, those based on quadratic error, on graph theory, hierarchical, and others. For the purposes of this study, segmentation techniques based on quadratic error were used since they are capable of handling large datasets that are frequently used in the context of transport, specifically the K-medoids algorithm [26], because it is one of the most robust against noise. A medoid may be defined as the object of a cluster whose average dissimilarity to all objects in the cluster is minimal. It is the point located the closest to the centre in the whole cluster.

The validity of a solution in a clustering problem is evaluated using validity indices. Following the classification proposed by Aldana-Bobadilla [27], these indices may be classified into three categories: external indices, which measure the extent to which cluster labels match externally-supplied class labels; internal indices, which measure the intrinsic information of each dataset, and relative indices, which are used to compare several different clustering solutions. For the purposes of this study, an internal index was chosen to measure the quality of the clusters in the first instance: the silhouette function [28]. This measures the consistency of the cluster based on a comparison of the tightness and separation of the elements of each segment generated and is computed by the following formula:

$$\mathrm{s(i)} = \begin{cases} 1 - \frac{\mathrm{a(i)}}{\mathrm{b(i)}}, if\ a(\mathrm{i}) < b(\mathrm{i}) \\ 0, if\ a(\mathrm{i}) = b(i) \\ \frac{\mathrm{b(i)}}{\mathrm{a(i)}} - 1, if\ b(\mathrm{i}) < a(\mathrm{i}) \end{cases} \tag{1}$$

In Formula (1), $a(i)$ is the average distance from object i to the other objects within the cluster and $b(i)$ is the smallest average distance from i to all the objects of each of the clusters to which i does not belong.

Considering different authors, such as Lathia [29], who indicated that it is not always the case that the optimal value resulting from applying various methods to define the cluster number is the most appropriate, two additional evaluations were carried out to analyse the result of the segmentations. The first—independent criterion—consisted in evaluating the new attributes based on the intrinsic characteristics of the data. To this end, mutual information based on Shannon entropy was used, as expressed in Formula 2, which, in this case, is defined by the difference in uncertainty about the number of passengers not incorporated and incorporating the new attribute:

$$I(\text{Class; Attribute}) = H(Class) + H(Attribute) - H(Class, Attribute) \tag{2}$$

The second evaluation method—dependent criterion—consisted in applying a data mining algorithm to ascertain the effect of the attributes and to select them. In our case, a neural network was trained using the indicated data to predict demand and the attribute was evaluated for the error generated with the test dataset. Neural networks for the estimation of demand are commonly used in the field of transport, due to their ability to process multidimensional data, their learning capacity and their predictive capacity [30]. Evidently, this second method is more costly in computational terms. The set of procedures involved is illustrated in Figure 2. This figure shows the main processes of this study, framed within the Data Preparation and Modelling phases:

- Data Preparation phase. Depending on the defined time granularity, creation of the datasets to be modelled from the origin-destination matrices.
- Data Modelling phase, divided in two main tasks: generation and evaluation of 2–*k* clusters, using an independent criterion and a dependent criterion.

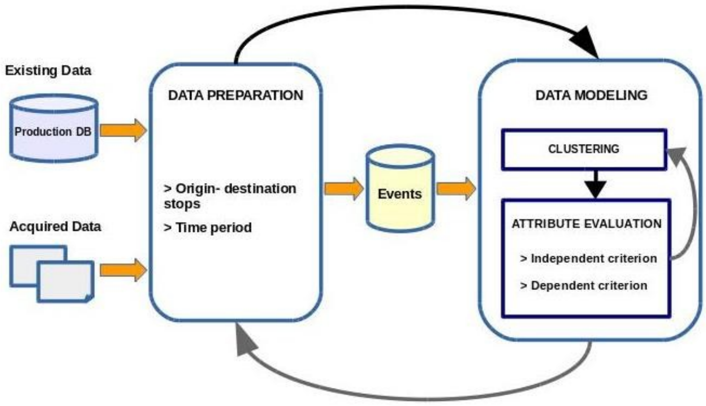


Figure 2. General scheme of processes in the evaluation of the new attributes.

4. Results

The methodology described was implemented in the intercity public transport company Global Salcai-Utinsa. It was used to study demand between three pairs of origin-destination stops on widely used routes by passengers of this company. Since demand was being analysed, six trips were studied for each pair of origin-destination stops, in both directions (outbound and inbound). The four stops were chosen according to different types of demand that should exist a priori depending on the socio-economic characteristics of the geographical areas in which they are located. These stops are described below:

- Stop identified with code 0. This stop corresponds to the main station in the city of Las Palmas de Gran Canaria, the capital of the island and the most populated municipality, which has the greatest number of public and private service centres.
- Stop identified with code 11. This stop corresponds to the main stop in the municipality of San Bartolomé de Tirajana. This is the biggest tourist municipality on the island of Gran Canaria; according to records, in 2016 it had 88,297 tourist beds [31].
- Stop identified with code 66. This stop corresponds to the main stop in the municipality of Santa Brígida. This municipality is a dormitory town for the capital of the island and its per-capita income is the highest on the island.
- Stop identified with code 99. This stop corresponds to Gran Canaria Airport. In 2016, this airport was used by 12,093,646 passengers, and is ranked fifth in the list of Spanish airports by the number of passengers [32].

From these stops, the following pairs of origin-destination stops were selected to analyse demand on the inbound and outbound trips:

- Trips between stops 0 and 66: trip from stop 0 to stop 66 (0–66) and trip from stop 66 to stop 0 (66–0).
- Trips between stops 0 and 11: trip from stop 0 to stop 11 (0–11) and trip from stop 11 to stop 0 (11–0).
- Trips between stops 99 and 11: trip from stop 99 to stop 11 (99–11) and trip from stop 11 to stop 99 (11–99).

With regard to the tools used, in the data preparation phase, Oracle was used for the database system and Pentaho for integration and visualisation. In the Modelling phase, the RStudio framework was used; more specifically, the Cluster [33], FSelector [34] and Neuralnet [35] modules. The data related to demand ($A_{W,D,H}$ in the dataset for classifying demand and N in the dataset for evaluating demand) were scaled in the processes of clustering and prediction with neural networks, using the maximum and minimum values for each of the routes. In the process of clustering, the metric used to calculate the dissimilarities between observations was the Euclidean distance, and the medoids were not initially specified. Nor were the weights of the neurons of the hidden layer initialised in the processes of creating the neural networks, where the differentiable error function used was the sum of squared errors.

4.1. Creation of the New Attribute

The results obtained in the creation phase of the new attributes are described below. As stated above, the structure of the data used is illustrated in Table 1. An origin–destination matrix was generated for each of the analysed routes, based on the direct or contactless card payment records of the passengers that made this trip, regardless of the bus line, between the first and final date of the analysis and during a specific time period. The period analysed was all of 2015, from 6:00 a.m. to 10:00 p.m. Table 3 shows the number of passengers on the trips analysed according to the means of payment used. From these initial specifications to define the period of analysis and considering the meaning of the data fields in Table 1 (dataset for classifying demand), the following values are used:

- P_o–P_d may have the following values: 0–66, 66–0, 0–11, 11–0, 99–11 and 11–99.
- W has a value of between 1 and 53, since 2015 had 53 weeks.
- D has a value of between 1 and 7.
- $A_{W,D,H}$ total number of passengers on day D, of week W, at time H from stop P_o to stop P_d. H_0 indicates the first hour of the day analysed (06:00) and H_f indicates the last hour of the day analysed (22:00).

Table 3. Number of passengers on each trip according to the means of payment used.

Trip O–D	Direct Payment Passengers	Card Payment Passengers
Trip 0–66	53,067	34,127
Trip 66–0	41,701	23,848
Trip 0–11	135,715	17,173
Trip 11–0	125,181	15,074
Trip 99–11	73,026	2086
Trip 11–99	76,214	2251

From the dataset for each trip, comprising the values indicated in Table 1 ($A_{W,D,H0}$, $A_{W,D,H1}$, ..., $A_{W,D,Hf}$), seven different clusters were created according to the number of segments considered, from 2 to 8, using the K-medoids algorithm. The consistency of each of the seven clusters was evaluated with the silhouette function.

Figure 3 shows the mean value for the silhouette function [26] obtained in each of the seven clusters for each of the analysed routes. The figure shows that the groupings of 2, 3 and 4 clusters are those that obtain higher consistency values. Owing to this result, only the new attributes resulting from these three cluster groups were evaluated; these new attributes were named *K2*, *K3* and *K4*.

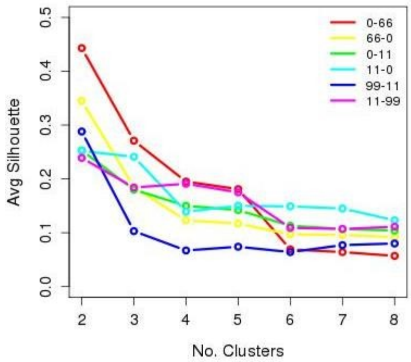
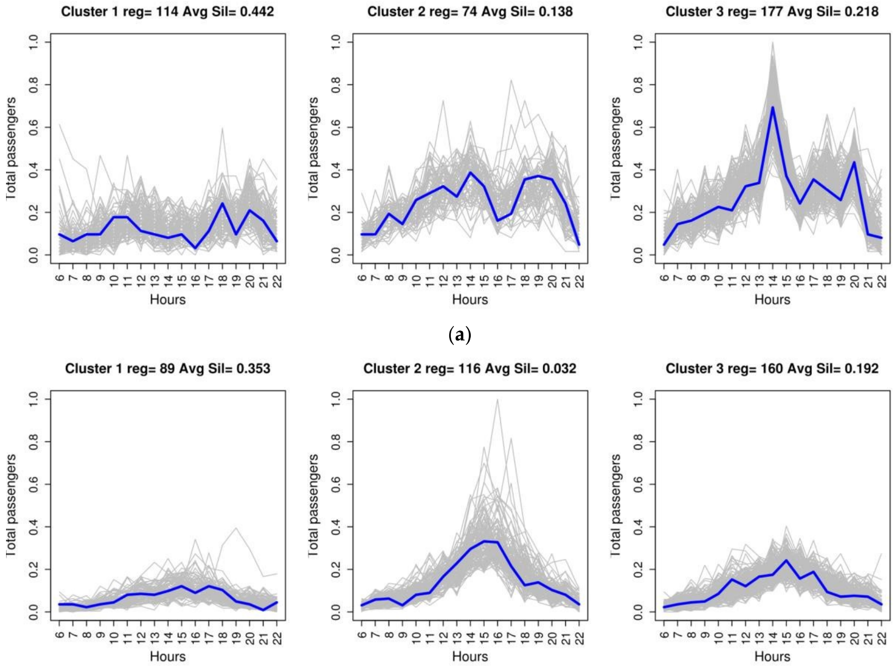


Figure 3. Result of the consistency analysis using the silhouette function. The horizontal axis represents the number of clusters considered in each clustering process. The average values obtained with the silhouette function are represented on the vertical axis.

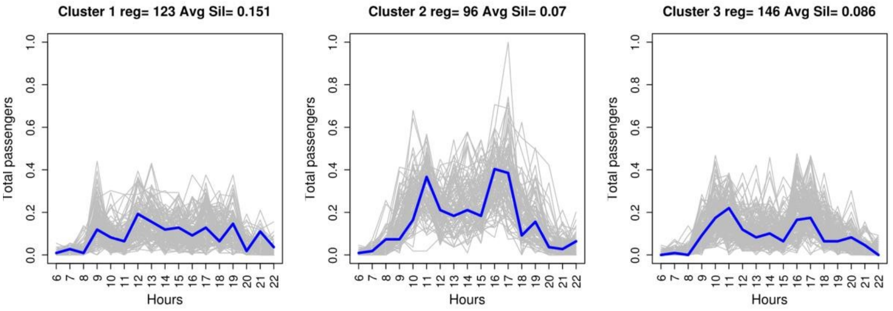
Figure 4 illustrates the clustering result with three clusters for three of the six routes analysed, with the representative medoid of each cluster in blue.



(**a**)

(**b**)

Figure 4. *Cont.*



(**c**)

Figure 4. Clustering of three of the trips analysed. The horizontal axis represents the time of the day, from 6:00 a.m. to 10:00 p.m. The vertical axis represents the number of passengers, normalised. In the upper part of the graph for each cluster, the number of data records for that cluster and the consistency value of the cluster obtained with the silhouette function are indicated: (**a**) Clusters 1, 2 and 3 of Trip 0–66; (**b**) Clusters 1, 2 and 3 of Trip 0–11; and (**c**) Clusters 1, 2 and 3 of Trip 99–11.

This task gave a relationship between the fields P_0, P_d, W and D with the new attributes, such that, in the next evaluation phase, the value of said attribute was assigned to the C field of the corresponding data record of Table 2.

4.2. Evaluation of the New Attributes.

The new attributes were evaluated according to two different criteria. The first applied criterion was the independent criterion, which consists of evaluating the information gain using the Shannon entropy function. To do this, the value of this gain was calculated for each of the attributes frequently used in planning (month, week of the year, day of the week and public holiday) and the new attributes (*K2*, *K3* and *K4*) with respect to the number of passengers class (attribute N of the data records from the learning and test datasets represented in Table 2).

Figure 5 shows the entropy values obtained for the analysed trips and the values obtained for each of the hours of greatest passenger numbers on each of them: 2:00 p.m.–3:00 p.m. on Trip 66–0; 7:00 a.m.–8:00 a.m. on Trip 66–0; 3:00 p.m.–4:00 p.m. on Trip 11–0; 10:00 a.m.–11:00 a.m. on Trip 0–11; 5:00 p.m.–6:00 p.m. on Trip 11–99 and 8:00 a.m.–9:00 a.m. on Trip 99–11. The graphs show that the new generated attributes individually obtain a high information gain value in comparison with the classic time-dependent attributes used as reference.

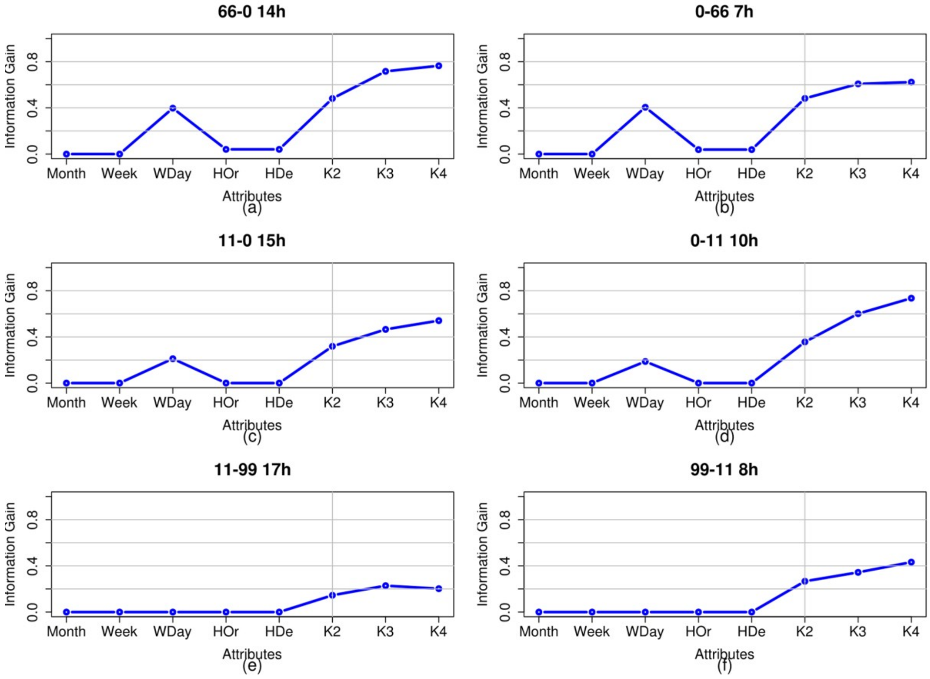


Figure 5. Shannon entropy values for the classic attributes and the new attributes against the number of passengers class for each of the trips analysed in the peak time period for each of them. The evaluated attributes are represented on the horizontal axis. The vertical axis represents the information gain value using the Shannon entropy function. At the top of each graph, the trip and the peak time of the trip are indicated. (**a**) entropy of Trip 0–66; (**b**) entropy of Trip 66–0 ; (**c**) entropy of Trip 0–11; (**d**) entropy of Trip 11–0; (**e**) entropy of Trip 99–11; (**f**) entropy of Trip 11–99.

As a second evaluation criterion, the effect of the new attributes was used to predict demand through neural networks. Specifically, "Resilient Backpropagation with Weight Backtracking" networks were used, due to their rate of convergence. The number of input neurons was determined by the total number of attributes used and the number of neurons in the hidden layer varied between one and five. In all cases, there is an output neuron with the estimated total number of passengers.

In this case, the network training and evaluation procedure was carried out using the cross-validation technique that provides an estimate with low bias [36], using the same records that correspond to the year of study (2015). Specifically, five folders were used. The observed error, *Eo*, in the prediction is the average value of the error obtained, *E*(*i*), for each of the events, *i*, and is calculated as the difference between the normalised values of the real value, *VR*(*i*), and the estimated value *VE*(*i*):

$$\mathrm{E(i)} = |\mathrm{VR(i)} - \mathrm{VE(i)}| \tag{3}$$

$$\mathrm{Eo} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{\mathrm{E(i)}}{\mathrm{n}} \times 100 \tag{4}$$

Figure 6 shows the errors observed when estimating demand with the month and day of the week attributes and when it was estimated adding each of the new generated attributes to them. As was done with the entropy evaluation, this estimate was made in the time period with the greatest number of travellers on each route analysed. In all cases, it can be seen that the incorporation of the new attribute reduces the error in the estimation; in the case of Trip 0–66, this is close to 40% for the network with five neurons in the hidden layer. Table 4 gives a summary of the aforementioned results, representing the mean observed error on each of the routes for the different neural network configurations used. The baseline was the prediction using only the month and day of the week–attributes that are frequently used in this type of prediction [37]. As can be seen when adding any of the new attributes, the error in the prediction was reduced, in some cases by up to 30%.

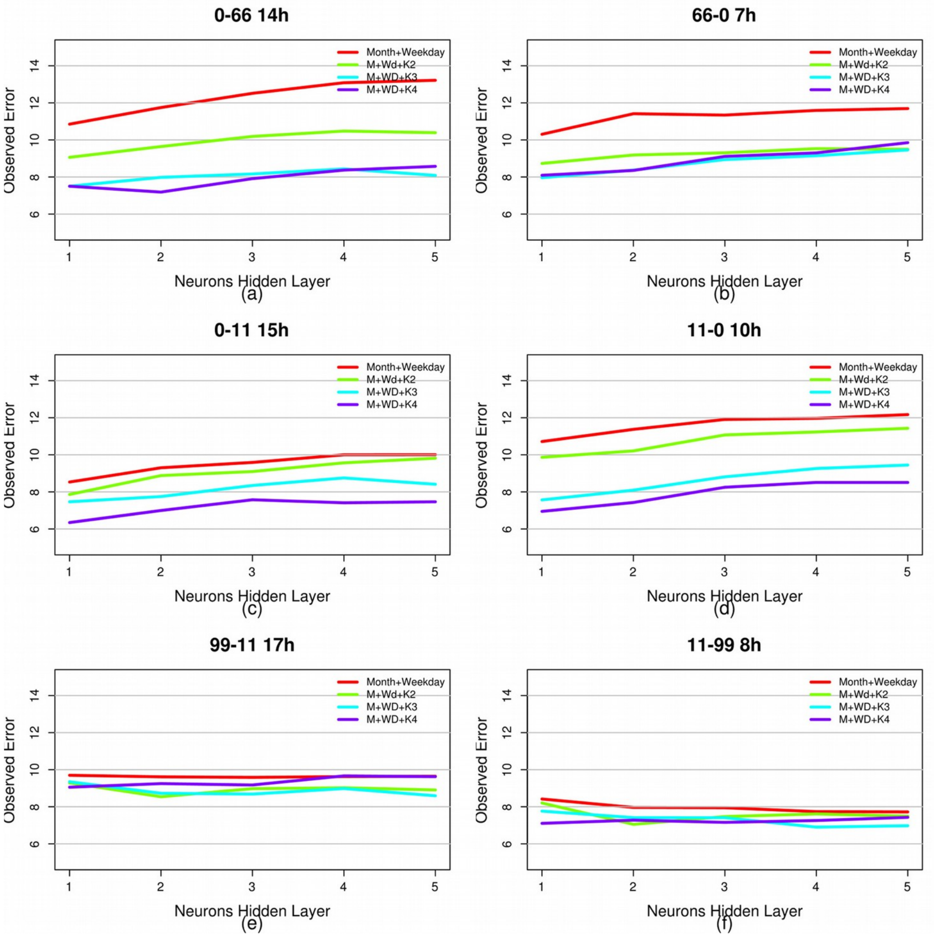


Figure 6. Observed error in the prediction of demand on each of the trips analysed in the peak periods of each one of them. The number of neurons in the hidden layer is represented on the horizontal axis. The vertical axis represents the observed error. At the top of each graph, the trip and peak time for which the prediction was made are indicated. (**a**) error on Trip 0–66; (**b**) error on Trip 66–0 ; (**c**) error on Trip 0–11; (**d**) error on Trip 11–0; (**e**) Error on Trip 99–11; (**f**) Error on Trip 11–99.

When choosing the number of clusters to be used to obtain the new attributes, two mistakes must be avoided: the first is that the number of clusters chosen is greater than the number of clusters that actually exist in the data, and the second that the number of clusters chosen is lower. Although it depends on the context of the problem to be solved, it is worse to commit the latter error since this would lead to a loss of information when the results are interpreted [38]. Considering the above and observing, firstly, that the best consistency value was obtained using two clusters and, secondly, that analysis of the clusters, in terms of the information provided and the error committed in the prediction, shows that better results are obtained with three and four clusters, it was decided that three clusters would be used to interpret the results in the following Discussion section.

Table 4. Mean observed error for the different configurations of the neural network (number of units in the hidden layer).

	Trip (Peak Period)					
Attributes	0–66 (14 h)	66–0 (7 h)	0–11 (15 h)	11–0 (10 h)	99–11 (17 h)	11–99 (8 h)
Month and Day of the week (baseline)	12.28	11.27	9.49	11.63	9.64	7.96
Month, Day of the week and *K*2	9.95	9.25	9.04	10.76	8.95	7.57
Month, Day of the week and *K*3	8.03	8.78	8.15	8.64	8.87	7.29
Month, Day of the week and *K*4	7.91	8.94	7.16	7.93	9.35	7.25

5. Discussion

Evaluation of the proposed new attributes show that they can provide information in addition to that provided by traditional time-dependent attributes. From the demand patterns, represented by the new attributes, more precise knowledge of demand can be obtained if we analyse each one of the obtained clusters. One way to analyse the results is to use contingency tables to represent the proportion of data from each of the clusters belonging to different periods of time, in order to obtain the types of time periods that significantly affect demand. Figure 7 shows, for the grouping of three clusters, two contingency tables for three of the six trips analysed, one with the months of the year (Figure 7a,c,e) and the second with the days of the week (Figure 7b,d,f).

It can be seen in Figure 7a that, in the case of Trip 0–66, Cluster 3, which corresponds to the highest demand profile (Figure 4a) is concentrated mainly in 10 months of the year (January–June and September–December). These periods correspond to school periods in which the use of public transport to go to schools pushes up the number of public transport users. For this same route, Cluster 2, which is associated with an average demand profile, is concentrated in the months of July and August, a holiday period in which travel to school decreases. Cluster 1, which corresponds to the lowest demand profile, has a uniform distribution over every month of the year and is concentrated on days of lower working activity (Saturday and Sunday) (Figure 7b).

For Trip 0–11, the interpretation of the new attributes indicates a demand that is different and more complex from the previous patterns. In this case, the cluster with the highest demand profile is Cluster 2, with data records concentrated mainly from January to March, and, in August, November and December (Figure 7c). It is followed by Cluster 3, which is mainly concentrated in the period from April to June. Taking into account the time slots in which demand peaks occur (Figure 4b) and the fact that most travellers use direct payment instead of card payment (Table 3), this behaviour can be interpreted as visitors returning from the city of Las Palmas de Gran Canaria to the tourist centre where they are staying (San Bartolomé), since Cluster 2 corresponds to the months of the high tourist season and Cluster 3 to the low season. This demand can also be explained by the overlapping of two types of high and low season seen in the tourist destination of San Bartolomé de Tirajana. In winter, visitors come mainly from the countries of northern and central Europe, while, in the summer months, domestic tourism predominates. As was the case with the previous trip, a specific demand profile for Saturdays and Sundays reappears, corresponding to Cluster 1 (Figure 7d).

For Trip 99–11, in the analysis with contingency tables of attribute *K*3, they show that the cluster with the highest demand profile is Cluster 2, which mostly corresponds to the months of January, February, March, November and December (Figure 7e). This behaviour can be interpreted as the flow of travellers arriving at the airport and travelling to their place of accommodation in the tourist high season, which corresponds to visitors from central and northern Europe for whom the use of public transport is a more common habit than it is for domestic tourists. In addition, the contingency table for the days of the week (Figure 7f) shows demand that is different from the two previously mentioned trips. In this case, the days of lowest passenger numbers do not come on the weekend; they are on

Tuesday and Thursday. This weekly demand can be interpreted as being related to flight scheduling by the airlines.

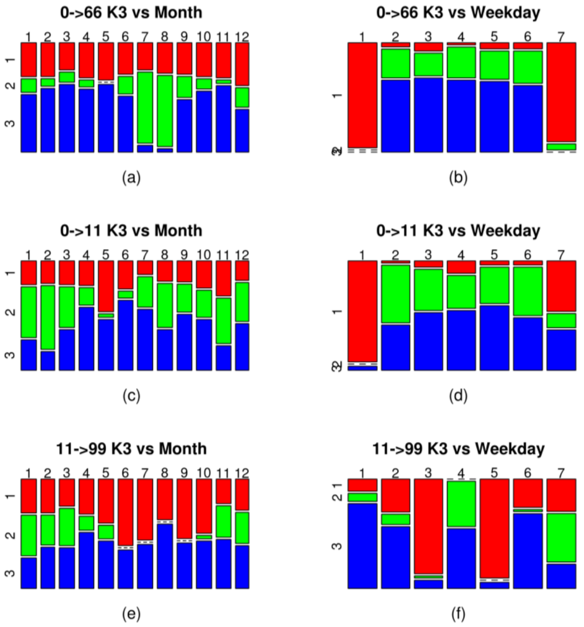


Figure 7. Graphic representation of the contingency table for the groupings of three clusters of three trips of the six that were analysed, according to the month of the year and the day of the week. The horizontal dimension is associated with a period of time (month in the tables on the left and day of the week in the tables on the right). The vertical dimension is associated with the clusters. (**a**) clusters of Trip 0–66 according to the month; (**b**) clusters of Trip 0–66 according to the day of the week; (**c**) clusters of Trip 0–11 according to the month; (**d**) clusters of Trip 0–11 according to the day of the week; (**e**) clusters of Trip 11–99 according to the month; (**f**) clusters of Trip 11–99 according to the day of the week.

Analysis of the medoids of the new attributes also makes it possible to infer information about demand, particularly when jointly studying the medoids of the trips between the pairs of stops under analysis. This analysis sheds light on which stop attracts the most passengers on that route depending on the time and makes it possible to infer what type of passenger makes the trip. Figure 8 shows the three medoids of Trips 0–11 and 11–0, to illustrate this aspect of the discussion. A significant common characteristic of the trips, pertaining to the profiles of greatest demand on these two trips (Figure 8b,c for Trip 0–11 and Figure 8e,f for Trip 11–0), is the fact that, for both trips, the time periods in which the peaks of greatest demand are produced in the two groups coincide. For Trip 0–11, from the main station in Las Palmas de Gran Canaria to the main stop in San Bartolomé, this peak time occurs between 2:00 p.m. and 4:00 p.m. For Trip 11–0, from the main stop in San Bartolomé to the main station in Las Palmas de Gran Canaria, this peak in demand occurs between 9:00 a.m. and 11:00 a.m. This behaviour leads to the conclusion that the typical passenger between these stops travels from stop 11 to 0 between 9:00 a.m. and 11:00 a.m., and returns between 2:00 p.m. and 4:00 p.m. It may also be concluded that this passenger is a visitor who is staying in the tourist town of San Bartolomé and is making a short visit, of between three and five hours, to the city of Las Palmas de Gran Canaria; the duration of the visit matches the visitor's time habits. This interpretation, based on the cluster medoids, is coherent and reinforces the interpretation of the contingency tables for this route: associating the periods in which the two patterns of greatest demand were produced with the island's two tourist seasons.

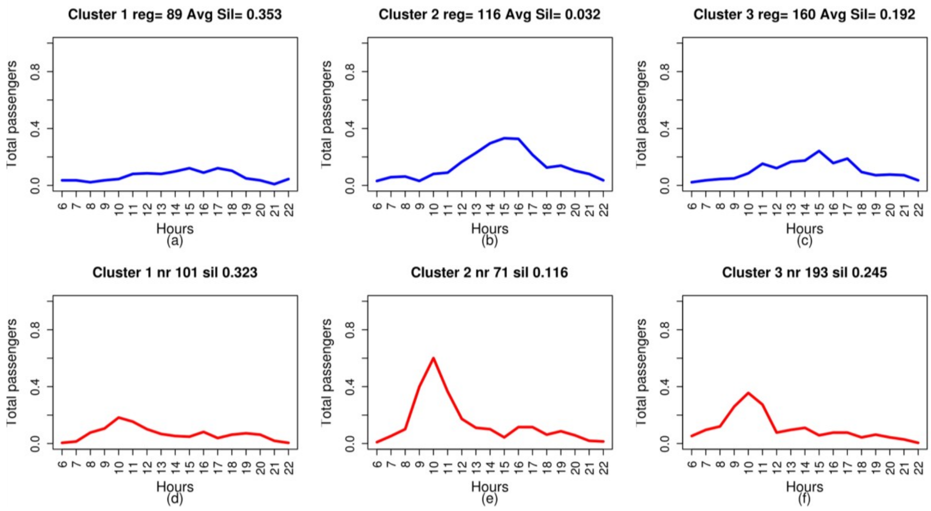


Figure 8. Medoids of three clusters of the two trips between stops 0 and 11. The medoids in blue are those obtained for the route between stop 0 and 11 and the medoids in red are those obtained for the route between stop 11 and 0. The horizontal axis represents the time of day, from 6:00 a.m. to 10:00 p.m. The vertical axis represents the number of passengers, normalised: (**a**) medoid of Cluster 1 on Trip 0–11; (**b**) medoid of Cluster 2 on Trip 0–11; (**c**) medoid of Cluster 3 on Trip 0–11; (**d**) medoid of Cluster 1 on Trip 11–0; (**e**) medoid of Cluster 2 on Trip 11–0; (**f**) medoid of Cluster 3 on Trip 11–0.

With regard to the prediction made using the new attributes, this was done using cross-validation since the main objective was to evaluate them and the method used to generate them. However, the results are promising since the incorporation of the new attributes reduces error when compared to traditional attributes, with the exception of two configurations of the network on Trip 99–11, where this error is equal.

6. Conclusions

This paper proposes a new type of time-dependent attribute to classify demand on public road transport systems. Moreover, the methodology that was followed has been described and implemented to facilitate the systematic generation of these attributes. The methodology is based on data mining.

From the results that were obtained from implementation in a real use case of demand analysis in a public passenger transport company, it was concluded that the proposed new attributes provide more information than the time-dependent attributes that have traditionally been used to design transport networks and to schedule public road transport systems. In addition, these new attributes enable us to classify demand over different time scales, taking into account different factors, and to obtain information on aspects such as the stops that attract the most passengers, type of traveller and reason for the trip.

The results have demonstrated the suitability of the methodology for the systematic acquisition of these new attributes. It is based on data mining and, more specifically, on clustering techniques, entropy analysis and neural networks. Finally, this methodology can be implemented in most intercity public transport companies, since it uses data that are usually found in companies' information systems, so it does not require external data sources.

Acknowledgments: The authors wish to express their gratitude to the Autoridad Única del Transporte de Gran Canaria (Transportation Agency of Gran Canaria) and to the public transport company Global Salcai-Utinsa S.A. They also declare that they have not received any funds to cover the publication costs of this article.

Author Contributions: C.R.G. coordinated the research tasks for this paper; all the authors contributed to the definition of the objectives; T.C. and G.P. were responsible for the data preparation phase and preparation of the required tools; all the authors participated in obtaining the new attributes; T.C., J.L.-N., A.Q.-A. and C.R.G. evaluated the new attributes; and all of the authors contributed to the preparation of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. International Energy Agent. Word Energy Outlook 2012. Available online: https://www.iea.org/publications/freepublications/publication/world-energy-outlook-2012.html (accessed on 13 December 2017).
2. World Health Organization. WHO Releases Country Estimates on Air Pollution Exposure and Health Impact. Available online: http://www.who.int/mediacentre/news/releases/2016/air-pollution-estimates/en/ (accessed on 13 December 2017).
3. European Commission. Road Safety: EU Reports Lowest Ever Number of Road Deaths and Takes First Step towards an Injuries Strategy. Available online: http://europa.eu/rapid/press-release_IP-13-236_en.htm (accessed on 24 January 2017).
4. Moreira-Matias, L.; Mendes-Moreira, J.; de Sousa, J.F.; Gama, J. Improving Mass Transit Operations by Using AVL-Based Systems: A Survey. *IEEE Trans. Intell. Trans. Syst.* **2015**, *16*, 1636–1653. [CrossRef]
5. Desaulniers, G.; Hickman, M.D. Chapter 2: Public Transit in Transportation. In *Handbooks in Operations Research and Management Science*; Barnhart, C., Laporte, G., Eds.; Elsevier: New York, NY, USA, 2007; Volume 14, pp. 69–127, ISBN 978-0-444-51346-5.
6. Yan, S.; Chen, H.L. A Scheduling Model and a Solution Algorithm for Inter-city Bus Carriers. *Transp. Res. Part A Policy Pract.* **2002**, *36*, 805–825. [CrossRef]
7. Yan, S.; Chi, C.; Tang, C. Inter-city Bus Routing and Timetable Setting Under Stochastic Demands. *Transp. Res. Part A Policy Pract.* **2006**, *40*, 572–586. [CrossRef]
8. Berbey-Alvarez, A.; Merchan, F.; Calvo-Poyo, F.J.; Caballero-George, R.J. A Fuzzy Logic-Based Approach for Estimation of Dwelling Times of Panama Metro Stations. *Entropy* **2015**, *17*, 2688–2705. [CrossRef]
9. Patnaik, J.; Chien, S.; Bladikas, A. Using Data Mining Techniques on APC Data to Develop Effective Bus Scheduling Plans. *J. Syst. Cybern. Inf.* **2006**, *4*, 86–90. [CrossRef]
10. Mendes-Moreira, J.; Moreira-Matias, L.; Gama, J.; Freire de Sousa, J. Validating the Coverage of Bus Schedules: A Machine Learning Approach. *Inf. Sci.* **2015**, *293*, 299–313. [CrossRef]
11. Khiari, J.; Moreira-Matias, L.; Cerqueira, V.; Cats, O. Automated Setting of Bus Schedule Coverage Using Unsupervised Machine Learning. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Auckland, New Zealand, 19–22 April 2016.
12. Furth, P.G.; Muller, T.H.J. Service Reliability and Optimal Running Time Schedules. *J. Transp. Res. Board* **2007**, *2034*, 55–61. [CrossRef]
13. Dessouky, M.; Hall, R.; Zhang, L.; Singh, A. Real-time Control of Buses for Schedule Coordination at a Terminal. *Transp. Res. Part A Policy Pract.* **2003**, *37*, 145–164. [CrossRef]
14. Hadas, Y.; Ceder, A. Optimal Coordination of Public-transit Vehicles Using Operational Tactics Examined by Simulation. *Transp. Res. Part C Emerg. Technol.* **2010**, *18*, 879–895. [CrossRef]
15. Saéz, D.; Cortés, C.; Milla, F.; Nuñez, A.; Tirachini, A.; Riquelme, M. Hybrid Predictive Control Strategy for a Public Transport System with Uncertain Demand. *Transportmetrica* **2012**, *8*, 61–86. [CrossRef]
16. Agard, B.; Partovi-Nia, V.; Trépanier, M. Assessing Public Transport Travel Behaviour from Smart Card Data with Advanced Data Mining Techniques. In Proceedings of the 13th World Conference on Transport Research, Rio de Janeiro, Brazil, 15–18 July 2013.
17. Lathia, N.; Capra, L. Mining Mobility Data to Minimise Travellers' Spending on Public Transport. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011.

18. Du, B.; Yang, Y.; Lv, W. Understand Group Travel Behaviors in an Urban Area Using Mobility Pattern Mining. In Proceedings of the IEEE 10th International Conference on Ubiquitous Intelligence and Computing and IEEE 10th International Conference on Autonomic and Trusted Computing, Vietri sul Mare, Italy, 18–21 December 2013.
19. Sun, F.; Pan, Y.; White, J.; Dubey, A. Real-time and Predictive Analytics for Smart Public Transportation Decision Support System. In Proceedings of the 2016 IEEE International Conference on Smart Computing, Saint Louis, CA, USA, 18–20 May 2016.
20. Yuan, J.; Zheng, Y.; Xing, X. Discovering Regions of Different Functions in a City Using Human Mobility and POIs. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012.
21. Blei, D.M. Probabilistic Topic Models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
22. Zhao, K.; Prasath, M.C.; Tarkoma, S. Automatic City Region Analysis for Urban Routing. In Proceedings of the IEEE International Conference on Data Mining workshop 2015, Atlantic City, NJ, USA, 14–17 November 2015.
23. Shearer, C. The CRISP-DM Model: The New Blueprint for Data Mining. *J. Data Wareh.* **2000**, *5*, 13–23.
24. Global Transporte Interurbano. Available online: http://www.globalsu.net/es/global_numeros.php (accessed on 13 December 2017).
25. Xu, R. Survey of Clustering Algorithms. *IEEE Trans. Neural Netw.* **2005**, *16*, 645–678. [CrossRef] [PubMed]
26. Kaufman, L.; Rousseeuw, P.J. Partitioning Around Medoids (Program PAM). In *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 1990; pp. 68–125. [CrossRef]
27. Aldana-Bobadilla, E.; Kuri-Morales, A. A Clustering Method Based on the Maximum Entropy Principle. *Entropy* **2015**, *17*, 151–180. [CrossRef]
28. Rousseeuw, P.J. Silhouettes: A graphical Aid to the Interpretation and validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
29. Lathia, N.; Smith, C.; Froehlich, J.; Capra, L. Individuals Among Commuters: Building Personalized Transport Information Services from Fare Collection Systems. *Perv. Mob. Comput.* **2013**, *9*, 643–664. [CrossRef]
30. Karlaftis, M.G.; Vlahogianni, E.I. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Trans. Res. Part C Emerg. Technol.* **2011**, *19*, 387–399. [CrossRef]
31. Instituto Canario de Estadística. Available online: http://www.gobiernodecanarias.org/istac/temas_estadisticos/sectorservicios/ (accessed on 13 December 2017).
32. AENA. Available online: http://www.aena.es/csee/Satellite?pagename=Estadisticas/Home (accessed on 13 December 2017).
33. Cluster Analysis Basics and Extensions. R Package Version 2.0.6. Available online: https://CRAN.R-project.org/package=cluster (accessed on 13 December 2017).
34. FSelector: Selecting Attributes. R Package Version 0.21. Available online: https://CRAN.R-project.org/package=FSelector (accessed on 13 December 2017).
35. neuralnet: Training of Neural Networks. R Package Version 1.33. Available online: https://CRAN.R-project.org/package=neuralnet (accessed on 13 December 2017).
36. Bradley, E. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *J. Am. Stat. Assoc.* **1983**, *78*, 316–331.
37. Celebi, D.; Bolat, B.; Bayraktar, D. Light Rail Passenger Demand Forecasting by Artificial Neural Networks. In Proceedings of the 2009 International Conference on Computers and Industrial Engineering, Troyes, France, 6–8 July 2009.
38. Milligan, G.W.; Cooper, M.C. An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika* **1985**, *50*, 159–179. [CrossRef]

2.2. System Proposal for Mass Transit Service Quality Control Based on GPS Data

sensors

MDPI

Article

System Proposal for Mass Transit Service Quality Control Based on GPS Data

Gabino Padrón, Teresa Cristóbal, Francisco Alayón, Alexis Quesada-Arencibia * and Carmelo R. García *

Institute for Cybernetics, Campus de Tafira, Las Palmas de Gran Canaria,
University of Las Palmas de Gran Canaria, Las Palmas 35017, Spain; gabino.padron@ulpgc.es (G.P.);
teresa.cristobalb@gmail.com (T.C.); francisco.alayon@ulpgc.es (F.A.)

* Correspondence: alexis.quesada@ulpgc.es (A.Q.-A.); ruben.garcia@ulpgc.es (C.R.G.);
Tel.: +34-928-458-651 (C.R.G.); Fax: +34-928-458-700 (C.R.G.)

Received: 21 April 2017; Accepted: 13 June 2017; Published: 16 June 2017

Abstract: Quality is an essential aspect of public transport. In the case of regular public passenger transport by road, punctuality and regularity are criteria used to assess quality of service. Calculating metrics related to these criteria continuously over time and comprehensively across the entire transport network requires the handling of large amounts of data. This article describes a system for continuously and comprehensively monitoring punctuality and regularity. The system uses location data acquired continuously in the vehicles and automatically transferred for analysis. These data are processed intelligently by elements that are commonly used by transport operators: GPS-based tracking system, onboard computer and wireless networks for mobile data communications. The system was tested on a transport company, for which we measured the punctuality of one of the routes that it operates; the results are presented in this article.

Keywords: intelligent transport systems; public transport systems; operation control; automatic vehicle location

1. Introduction

In modern societies, mobility plays a significant role in quality of life and is an aspect that has a considerable influence on the socioeconomic development of people. Because of this, we are currently witnessing a large-scale use of transport systems that is giving rise to problems such as environmental degradation, traffic congestion and an increased risk of road accidents. Different local, national and international agencies are constantly producing reports that highlight both the importance of transport systems and the need to address the problems inherent therein. It is estimated that over 40% of the world population spends at least an hour a day travelling by road [1]; the relationship between the rising number of people with heart, respiratory and brain diseases and high levels of pollution has been proven by medical studies [2]; the World Health Organisation [3] estimates that around 3 million people die annually due to outdoor air pollution; and finally, according to statistics, every year 250,000 people in the European Union [4] are seriously injured as a result of traffic accidents, with road fatalities at about 10% of that figure. Transport agencies are tackling these mobility-related problems mainly by, firstly, promulgating traffic rules and regulations that benefit transport and, secondly, using technology to develop more efficient transport systems. A highlight of the latter approach is the use of intelligent transport systems (ITS) to develop public transport systems that are more efficient, environmentally friendly and attract citizens away from the use of private transport. The key element in making a public transport system attractive to citizens is quality, an aspect that means responding to mobility needs and running on schedule. There is general consensus on the aspects that influence public transport users' perception of service quality. These are, mainly, punctuality, real-time availability of information on timetables including incidents affecting the service, and shorter waiting times at stops. There are even recommendations and standards, and legislation such as that which exists in the European Union [5], which sets out the parameters to be considered when assessing the quality of service provided by a public transport operator, which include timetable adherence. Because public road passenger transport is affected by external variables such as traffic and weather conditions, user demand, etc., comprehensive quality control in a transport network requires continuous assessment. And this requires elements from the transport network infrastructure, such as sensors, and computer and communications systems, to obtain the required data, which are on a massive scale in the case of medium to large transport networks.

This article presents a system for continuously and systematically evaluating the service quality of a public road passenger transport operator. This systematic evaluation is performed using vehicle tracking systems and mobile communications infrastructures used by transport operators, the set-up of which does not require any specific deployment of new devices. In addition, the data provided by the proposed system are integrated into the data model used by the transport operator, thus allowing detailed and continuous analysis of important aspects such as the punctuality or regularity of the services provided to users.

In addition to this introductory section, this article is organised into seven sections. The second section reviews related studies, focusing on the criteria, parameters and systems used to measure the quality of service in public transport by road. The third section explains a number of preliminary ideas and concepts to be used in the rest of the article: a general description of the positioning system used by the system (GPS), and a presentation of the conceptual and formal model of the entities involved in evaluating quality of service, as well as the criteria and parameters used. Having reviewed related studies, introduced the proposed system and formalised the problem, in the fourth section we explain the challenges that must be tackled in quality control of a public road passenger transport service. The fifth section describes the proposed system, looking specifically at its operating principles and its constituent components; the system was tested under real conditions in a local transport company on the island of Gran Canaria. In the sixth section the results obtained when evaluating the quality of service are discussed, based on the criterion of punctuality on one of the company's routes for a period of one year. Finally, the seventh section contains the main conclusions.

2. Related Studies

In this section we shall review the studies published about quality of service in regular road passenger transport, with particular reference to publications that deal with criteria and metrics associated with, and systems used to obtain the data required for, quality control. Peeks et al. [6] ranked the factors that influence public transport users' perception of quality according to different levels of importance. On the first level is safety and reliability, on the second level, the speed of the journey, and on the third level, convenience, comfort and finally, the experience. According to Van Oort [7], three factors have a particularly negative effect on the quality of public transport: unforeseen increases in waiting times for travellers, the time spent by the traveller in situations of overcrowding due to an overloaded public transport network, and delays in arrival times at destinations due to the variability of travel times. From the point of view of the traveller, the first two factors influence the feeling of comfort offered by the public transport service and the third factor influences their convenience-based decision on whether to use public or private transport. Moreira-Matias et al. [8] carried out an exhaustive review of the various methods that have been used to improve and optimise public passenger transport, which included the evaluation of adherence to public transport timetables. Timetable adherence means arriving at stops and stations on time, a key aspect in the user's perception of public transport service quality. It is therefore essential that timetables are created using techniques that schedule departure and arrival times as accurately as possible. In this vein, the problems in predicting public transport bus arrival times at stops have been addressed using different techniques. For example, Yu et al. [9] used Kalman filters, Chang et al. [10]

proposed a solution based on non-parametric regressions, and Jeong et al. [11] used neural networks. Once the route timetables have been established, various criteria and associated metrics have been used to evaluate adherence to those timetables. Turnquist [12] identified two strategies for planning regular road passenger transport routes: the first consists in planning routes by frequency, and the second, by timetable. The first planning strategy is used for urban transport and the criterion applied is regularity. The second planning strategy is used for long-distance transport and the criterion applied is punctuality. A starting point of reference is the study by Polus [13], in which he proposed the following indicators for arterial routes: overall travel time, congestion index, overall travel speed and overall delay. Subsequently, and making use of technological advances in onboard systems that enable a larger volume of data to be recorded, timetable adherence has been analysed at a far greater level of detail in the transport network: at a single stop, or on a given section of the route. Nakanishi [14], Strahman et al. [15] and Barabino et al. [16] proposed using the following parameters: on-time performance (OTP), run time variation (RTV), headway variation (HV) and excess waiting time (EWT). The first two are used to measure punctuality and the last two, to measure regularity. Lin et al. [17] proposed another indicator, called adjacent site punctuality rate. To obtain all these indicators it is necessary to gather data on vehicle location and the times that the vehicles pass through each of the control points on the route. Continuous monitoring of all the stops on the route gives rise to the challenge of managing a massive amount of data, more so when the transport network contains a considerable number of stops, which is the case of medium to large metropolitan areas.

According to Furth et al. [18], the data used to check and improve public passenger transport planning may be obtained through three different data sources: surveys of public transport users, and automatic passenger counter (APC) and automatic vehicle location (AVL) systems. The AVL systems provide the data source for vehicle location and times of arrival and departure at stops. Several positioning technologies are used by AVL systems. According to Riter et al. [19], they can be broken down into three categories: radiolocation systems that establish position with radio signal triangulation techniques, estimation of the direction and distance travelled (dead reckoning) or proximity detection systems. Currently, radiolocation systems are the most used, particularly the Global Positioning System (GPS) [20], although proximity detection systems are also used, such as the system proposed by Zhou et al. [21], who used a Radio Frequency Identification (RFID) system to detect the presence of the vehicle at a stop in order to assist travellers with special needs. Examples of AVL system proposals using GPS have been put forward by Mazlooumi et al. [22], who used GPS data to analyse public transport travel time variability; Zhao et al. [23], who used GPS data to determine optimal slack time for planning schedule-based services; Derevitskiy et al. [24] studied traffic conditions in the transport network using GPS data; Cortes et al. [25] used the same data to estimate bus speed; García-Castro et al. [26] used GPS data from private vehicles to calibrate the input data used in different techniques for modelling road traffic and the effects that it causes. GPS is a case of a global navigation satellite system (GNSS), which has become popular due to its features and the low cost of GPS receivers. In recent years navigation systems have been used that are able to use the signal from several different constellations of satellites to obtain more precise position measurements than those obtained using only signals from GPS satellites; along these lines, Dabove et al. [27] studied the use of the GLONASS constellation.

The system described in this article evaluates adherence to scheduled passing times on a public road passenger transport network. The system autonomously and continuously records all the events that describe the activity carried out by the transport vehicle (bus) by referencing them in space and time with a level of granularity that enables reliability evaluations using the various indicators mentioned above. The purpose of this may be to adapt to different modes of transport, such as long-distance transport routes on which timetable adherence at stops is a key element in the quality of service, or urban transport, where quality is associated with vehicle frequency. Our proposed system uses an AVL system that periodically records the vehicle position data every time the GPS reading is taken, integrating these data into the model used to represent the transport network, the planning of services and the activities carried out by the transport company.

3. Preliminary

3.1. Global Positioning System (GPS)

The positioning system used in our proposal is GPS. This system allows the location of mobile objects (airplanes, ships, cars, etc.) to be determined, thus enabling their route to be monitored. To provide geolocation information, GPS uses a constellation of satellites that emit a signal that is received and processed by the receivers built into mobile devices. When a GPS receiver is able to receive signals from at least four different satellites, a three-dimensional geolocation is obtained (latitude, longitude and altitude). In addition to position data, the GPS receiver also provides the time that the data were received; this information comes from the atomic clock on the GPS satellite, which is encoded and sent with the signal emitted by the satellite. The receiver also provides data indicating how many satellites were used to obtain the data (data quality) or the time elapsed since they were obtained (data age). In a conventional GPS receiver, position data errors vary depending on various factors, with errors of up to 100 m in the early days of civilian use, when the intentional degradation named Selective Availability (SA) was active [28]. There are positioning systems based on GPS that provide very precise measurements, such as Differential GPS, which achieves accuracy of less than 2 m. The errors in GPS data are due to different causes. A first source of error is found in the GPS system itself: in the errors that occur in the satellite orbits, caused by drift from true time by the atomic clocks on the satellites and by the geometric configurations of the satellites "visible" to the GPS receiver. Since its inception, GPS performance has gradually improved; Table 1 presents some performance metrics for the system from a study conducted in 2013 by the University of Texas, Austin [29]. Among other performance metrics analysed by this study, this table shows the metrics related to Signal-in-Space (SIS), User Range Error (URE), Accuracy, Position Service Availability, and Position Accuracy. AOD refers to Age of Data, which is provided for each GPS reading and indicates the time elapsed since the last position was calculated.

Table 1. A selection of GPS performance metrics.

Metric	Value
SIS URE Accuracy	$\leq$7.8 m 95% Global average URE during normal operations over all AODs $\leq$6.0 m 95% Global average URE during normal operations at zero AOD $\leq$12.8 m 95% Global average URE during normal operations at any AOD $\leq$30 m 99.94% Global average URE during normal operations $\leq$30 m 99.79% Worst-case single point average URE during normal operations
Position Service Availability	$\geq$99% Horizontal, average location $\geq$99% Vertical, average location $\geq$90% Horizontal, worst-case location $\geq$90% Vertical, worst-case location
Position Accuracy	$\leq$9 m 95% Horizontal, global average $\leq$15 m 95% Vertical, global average $\leq$17 m 95% Horizontal, worst site $\leq$37 m 95% Vertical, worst site

A second cause of error is the propagation medium, i.e., refraction in the ionosphere and troposphere, and errors produced by the same signal reaching the receiver by two or more different paths because it has bounced off buildings or natural elements; this error is called multipath error. A final cause of error in position data is due to errors in the GPS receivers. Our proposed system uses conventional GPS receivers installed in vehicles, and the accuracy of the data that they provide depends on the satellite constellation that they are able to view at the time that the position and geometry of the constellation is calculated. In an ideal scenario, with constellations of more than 7 satellites an accuracy of less than 2 m can be obtained. Given the different sources of error in the GPS coordinates, in the case of public transport vehicles, multipath errors are particularly relevant, especially when the vehicles run through urban areas. Considering the above explanation of GPS, the user accuracy depends on a combination of satellite geometry, URE, and local factors such as signal blockage, atmospheric conditions, and receiver design features/quality, adopting our system a maximum possible error in position data of 25 m, taking into account that the SIS URE Accuracy is behind 13 m during normal operations at any AOD and the metioned external factors (atmospheric conditions, multipath, technical manufacturer specifications, ...); henceforth this maximum error will be represented by E. Figure 1 is a map illustrating the position data captured by public transport buses that follow a route in an urban area. The blue dots represent the position data, and we can see that some of these coordinates stray from the road because of GPS data errors.

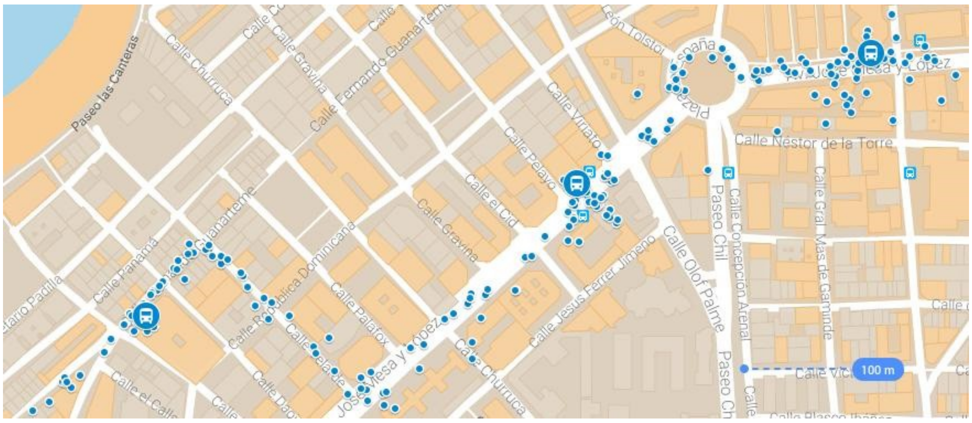


Figure 1. Examples of GPS location data for public transport buses on a fixed route. The positions are represented by the blue dots.

It is often the case that, once the vehicle's GPS geolocation has been obtained, it has to be positioned on the transport network. This requires a geographic database that contains the positions of the different points on the transport network. To locate a vehicle on the network, the distance between its GPS position and the position of the different points on the transport network is used. This distance can be calculated using different formulae, all of which result in errors since the Earth is not a perfect shape; it is neither a perfect sphere nor a perfect ellipsoid. In general, the distance formula with the lowest error is the haversine formula [30], and is therefore used by the proposed system. Given two points, $p1$ and $p2$, with coordinates expressed in radians (lat_1, lon_1) and (lat_2, lon_2), the value of this distance is calculated using the following expression:

$$Dist(p1, p2) = 2 \times R \times \arcsin(\sqrt{A}),$$

where R is the radius of the Earth at the equator and A, a factor calculated as follows:

$$A = \sin^2\left(\frac{\Delta lat}{2}\right) + \cos(lat_1) \times \cos(lat_2) \times \sin^2\left(\frac{\Delta lon}{2}\right),$$

$$\Delta lon = lon_2 - lon_1,$$

$$\Delta lat = lat_2 - lat_1.$$

This distance function produces an error, owing to the fact that the Earth is not a perfect sphere and the radius of curvature is not always the same. The maximum error occurs when the points are located at the greatest possible distance, i.e., in the antipodes—a distance of approximately 20,000 km—the maximum error being 2 km.

3.2. Conceptual and Formal Model of the Public Transport Network and Activity

This section will describe all the concepts involved in evaluating service reliability. These concepts are derived from international standards for conceptual transport data models; specifically, Transmodel [31], a European reference data model for public transport information. The first concept that we took from this conceptual model is that of transport network. According to this model, a transport network is made up of all the entities that represent the places, roads and routes necessary to spatially describe vehicle operations. One such entity is points on the transport network. These points may represent places where various actions are performed; for example: point where passengers board or alight from the vehicle, timetable control points, points of interest, etc. Each of these activities is associated with a classification, and a point may belong to more than one classification. For example, it may be both a point where passengers board and alight and a timetable control point. The points of particular interest for the purposes of our system are those points where passengers board or alight from the vehicle, which will be called, generically, stops. Examples of these points are platforms at stations or bus stops. In formal terms, the transport network may be modelled using a directed graph $G = (N, A)$, where N is the set of nodes, in which each node $p \in N$ represents a point of the transport network. A is the set of directed arcs, wherein each arc $a \in A$ represents the section of the route connecting two nodes $p, k \in N$, where p is the start node of the arc and k the end node of the arc. A route on the transport network is a route that runs systematically, starting at a node that is a stop on the route, the start node, and ending at another stop, the end node. The set of routes on the transport network is represented by R and each route on the network by r, $r \in R$. A route r is an ordered sequence of arcs such that $a \in A$, thus establishing the passing order for each node on the route: $r = (a_1, a_2, ..., a_\mathrm{n})$. The subscript of each arc of the sequence indicates the position in the ordered sequence, such that the start node on the first arc of the sequence, a_1, is the start node of the route and the end node of the last arc of the sequence, a_n, is the end node of the route. Given a route r, defined by a sequence of n arcs, the number of nodes involved in the definition of a line will be $n + 1$, of which at least two must be a stop. To evaluate the punctuality of route r, stop nodes are required. This subset of nodes of route r will be called route r set of stops, and will be represented by P_r; they are an ordered set of stop nodes, in which the order of each node in the set P_r matches the order in which they are visited on route r. A public transport operator's schedule, SP, consists of a set $SP = \{S_i\}$, where each element S_i is called a scheduling unit. Each scheduling unit, S_i, is assigned a vehicle resource, a driver resource and a set of operations to be carried out at a scheduled date and time. If the schedule is not met, this means an inefficient use of the allocated resources. The operations included in a S_i scheduling unit may be of different types, but for the purposes of monitoring quality of service, line service operations are of greatest interest. A line service is defined as the completion of route r at a scheduled date and time. The set of line service operations on route r, $L_r = \{l_i\}$, is the set that comprises all line service operations of route r established in SP; moreover, to represent only the line services on route r to be carried out during a period of time, T, the notation $L_{r,T}$ is used. For each line service, l_i, passing times through each of the nodes of set P_r are scheduled, and that will be the information advertised to transport users. On line service l_i on route r with n stops, the timetabled passing times of said line service is an ordered cumulative sequence of times specifying what time, t, the vehicle is scheduled to pass each of the stops on the route, i.e., $TP_{li} = \{t_1, t_2,..., t_\mathrm{n}\}$ where t_1 indicates at what time the vehicle is scheduled to start the line service, t_2 the arrival time at the second stop, and so on until the last time, t_n, which indicates the scheduled time at which the vehicle will reach the last stop on the route. The passing times of line service l_i on route r vary and depend on various factors, some of which are related to transportation issues and some of which are not. More specifically, when scheduling passing times at the various stops on a line service, it is necessary to consider the time required to complete each of the arcs that define the route and the time needed by passengers to board or alight from the vehicle at each stop on the route. The time spent completing the different arcs of the route depends on the vehicle speed, which in turn depends on variable factors such as weather conditions when the line service is running or the traffic status, which in turn is a factor dependent on the time of day the route is operated or the day type, e.g., working day, weekend or holiday, and even on the time of year. To draw this paragraph on conceptual models to a close, we would like to mention that they are implemented through a database that contains all the entities and the relationships between the entities representing the transport network, the deployed resources and the records reflecting all the facets of public transport operations. This database is called the transport database (TDB).

As mentioned in the previous section on related studies, there are different criteria for assessing adherence to the timetabled passing times on the line services included in *SP*. To decide on what criteria to use, the type of line service must be taken into account. For line services scheduled by frequency, typical of urban transport, the criterion used is the regularity with which the service is provided. However, for line services scheduled by timetable, typical of intercity transport, the criterion used is punctuality. As was stated in the previous section, to evaluate the punctuality of the line service, the OTP and RTV metrics are used. RTV is calculated using a basic metric called Arrival Delay (AD) obtained for each stop, p, of a line service on route r, $AD_{li,p}$. AD is defined as the difference between the observed passing time through stop p, $OT_{li,p}$, and the scheduled passing time through that stop, $ST_{li,p}$, when operating line service l_i. Considering that the passing times for each stop are measured with respect to the time at which the line service began, this metric is calculated using the following expression:

$$AD_{li,p} = OT_{li,p} - ST_{li,p}. \tag{1}$$

The RTV metric is calculated using the following expression:

$$RTV_{li} = n^{-1} \times \sum_{p=0}^{n} \frac{\left|OT_{li,p} - ST_{li,p}\right|}{OT_{li,p}} \tag{2}$$

As stated in the previous section, for line services operating by frequency, the criterion used is regularity. To evaluate regularity, the HV and EWT metrics are used. To calculate these metrics, first the Headway Ratio (HR) parameter has to be calculated using the following expression:

$$HR^{p}_{li,k} = \left(\frac{H^{p}_{li,k}}{f^{p}_{li,k}}\right) \times 100 \tag{3}$$

where $f^{p}_{li,k}$ is thescheduled frequency between two line services l_i, l_{i+1} at stop p on route r and $H^{p}_{li,k}$the frequency observed between both line services at stop p on route r. The factor 100 represents a perfect match between observed and scheduled time. If for each pair of possible values for service lines l_i, l_{i+1}, the average and standard deviation of HR at each stop p on route r is calculated, then the HV metric is obtained as follows:

$$HV^{p} = \frac{\sigma^{p}_{HR}}{\mu^{p}_{HR}} \tag{4}$$

The EWT metric is an estimate of excess user waiting time due to failure to meet the scheduled frequency of two line services on route r. Its value, at stop p on route r, is obtained from the value HV^{p} using the following expression:

$$EWT^{p} = \frac{(\sigma^{p}_{HR})^{2}}{2x\mu^{p}_{HR}} \tag{5}$$

4. The Service Quality Control Problem

As discussed above, quality of service on regular passenger transport by road is evaluated by applying criteria which vary depending on whether line services are planned by frequency or by passing times. In the first case, the criterion used is regularity and the associated metrics to evaluate regularity are HV and EWT. In the case of line services planned by timetabled passing times, the criterion used is punctuality and the associated metrics are AD and RTV. From the mathematical expressions for calculating each of these metrics we may deduce that in order to obtain the values it is necessary to know the time at which the vehicle arrives and stops—when it stops to pick up or set down passengers—or the time at which the vehicle passes the stop—when it does not stop because there are no passengers to pick up or set down at the stop. To obtain these data it is necessary to know the position of the vehicle, and from this position to locate the vehicle on the transport network, an action performed by the vehicle's positioning system. When locating the vehicle on the transport network, if the position is provided by GPS, it should be borne in mind that sometimes the location data are erroneous or may not be 100% accurate, depending on the data error caused by the various factors described in the previous section. When the vehicle passes a stop on the route without stopping, the time of passing is the moment at which it is nearest the stop; the distance between the stop and this point is defined by a preset threshold. By contrast, when the vehicle stops to pick up or drop off passengers, the arrival time at the stop is the moment at which the GPS data indicate that the vehicle is stationary and its distance in relation to the stop is below a defined threshold. When defining this threshold, the aforementioned GPS data error must be taken into account, as well as the variable length of the bus stop parking area; in the case of the transport network of this particular study this distance varied between 12 and 50 m. If vehicles only stopped at bus stops on their route, the arrival time at each stop could be obtained with a high degree of reliability, regardless of errors in GPS data. However, in reality no bus only stops at bus stops; they also stop because of traffic or because of the traffic signals along the route. The problem is how to distinguish the position of a bus that has stopped because of a traffic signal from the position of the same vehicle on the same route that has stopped at a bus stop when the position of the traffic signal and the stop are very close. This is a very common situation on urban bus routes.

Another challenge that must be addressed when studying the regularity or punctuality of line services on public transport is the reliability of the results. To achieve a high degree of reliability it is necessary to work with integral data sets, i.e., complete and correct sets of data. In the context of public transport, potential sources of error in the data are: failure in position data or in the devices that process the data, occasional errors due to noise or data that, although correct, have been obtained for routes that do not match the corresponding route plan. It is therefore necessary to carry out a filtering process to ensure data integrity when analysing the regularity or punctuality of line services.

A final challenge when conducting a comprehensive and ongoing evaluation of the punctuality or regularity of public transport services is the handling of the amount of data required, which in the case of medium to large transport networks is massive. Therefore, intelligent data handling is required.

5. System Overview

The proposed system is designed to continuously and comprehensively evaluate the quality of service provided by a regular road passenger transport operator. This evaluation is conducted using metrics commonly used by operators and transport agencies, measuring punctuality—e.g., with the AD and RTV parameters—or measuring regularity—e.g., with the HV or EWT parameters. Continuous evaluation of quality of service means that it may be performed at any given time, thus permitting its evolution to be analysed in different calendar periods (working days, holidays, weekends, time of year, etc.). Comprehensive evaluation means that the metrics can be obtained with different levels of granularity, i.e., for a set of routes, for specific routes, or for a set of individual stops.

The system primarily uses four resources that are usually employed in regular public passenger transport by road: GPS vehicle tracking system, the computing and storage systems used in vehicles to record and monitor the activities carried out on board, the communications system used to transfer data between the head office and the vehicles, and the transport database (TDB). The system consists of two modules: the AVL Module (AVLM) deployed on the vehicles and the Service Quality Control Module (SQCM) deployed at the operator's or transport agency's data processing centre. The AVLM continuously records and automatically transfers the vehicle's position. The SQCM evaluates the quality of service by calculating the punctuality and regularity metrics for the planned services. Figure 2 provides an overview of the system.

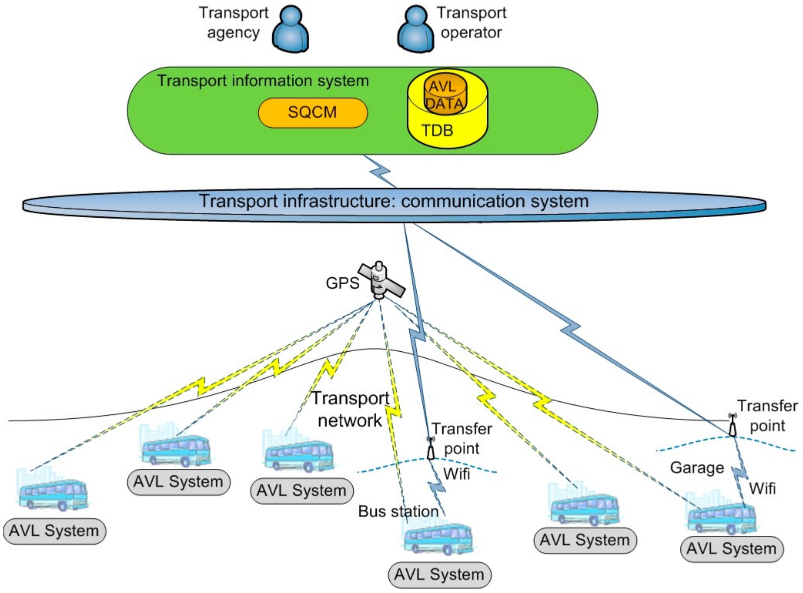


Figure 2. System Overview

5.1. The AVLM

The mission of this module is to obtain the basic data required to evaluate the punctuality or regularity of the line services performed by the vehicle. These data are obtained from the GPS vehicle tracking system. The data recorded are geodetic position, speed, data quality and the time at which the data were obtained. Table 2 shows the recording structure used to store each GPS location reading, specifying the meaning and size of each field. Henceforth these data will be referred to as RAVL.

The RAVL_{Age} and RAVL_{Sou} fields indicate the quality of the GPS data; data with a value of 2 in both fields indicates high quality, any data with a value of 0 in either of these fields indicate that the data are unreliable. A value of 1 in the Age field indicates that the data were obtained more than 10 s earlier. A value of 1 or 2 in the Source field indicates that it has been possible to obtain the latitude and longitude coordinates—i.e., two-dimensional location data—which occurs when the vehicle's GPS receiver was unable to obtain the signal from at least four GPS satellites.

The AVL module is run on the onboard computer typically used by transport companies to record events during vehicle service. This computer has a display, keyboard, cable communication interfaces to connect up the other devices on the vehicle (card readers, vehicle information panels, sensors, etc.), and wireless communication interfaces for data communications: 3G for long-distance data communications and WiFi for local communications. To transfer location data this module uses WiFi infrastructure.

Because of the potentially massive number of RAVL records—as the position of each vehicle is recorded continuously—the limited storage capacity of the onboard computer and the limited bandwidth of the WiFi communications, the AVLM handles data intelligently to minimise the following aspects:

- The number of AVL records required for quality control.
- The time AVL records are stored on the vehicle.
- Errors in data communications.

Table 2. Description of the RAVL data structure.

Field	Description (Size)
$\mathrm{RAVL}_{\mathrm{Veh}}$	Vehicle identifier. (2 bytes)
$\mathrm{RAVL}_{\mathrm{T}}$	GPS Time of day expressed in Coordinated Universal Time (UTC). (4 bytes)
$\mathrm{RAVL}_{\mathrm{Lat}}$	Latitude of the vehicle expressed in degrees. (4 bytes)
$\mathrm{RAVL}_{\mathrm{Long}}$	Longitude of the vehicle expressed in degrees. (4 bytes)
$\mathrm{RAVL}_{\mathrm{Alt}}$	Altitude of the vehicle position expressed in feet. (4 bytes)
$\mathrm{RAVL}_{\mathrm{Vel}}$	Horizontal speed of the vehicle expressed in miles per hour. (2 bytes)
$\mathrm{RAVL}_{\mathrm{Age}}$	Age of the GPS data: 0: data not available 1: >10 s 2: <10 s (1 byte)
$\mathrm{RAVL}_{\mathrm{Sou}}$	Type of position measurement: 0: data not available 1: 2D GPS 2: 3D GPS (1 byte)

To minimise the number of AVL records, this module uses a variable sampling period, which depends on the time unit used in planning timetables for each line service. For example, for long-distance line services the smallest unit of time is a minute. In addition, vehicle position sampling is only performed on a scheduled service, since when the vehicle is not in service or is performing an unplanned operation, no control is required.

To minimise the time that RAVL records are stored on the onboard computer, they may be transferred at different points of the transport network through which vehicles usually pass several times a day; for example, stations or garages.

To minimise errors in RAVL file transfers, caused by the fact that the vehicles are in motion and therefore WiFi connections are intermittent, the AVLM connects to WiFi by selecting the most suitable place and time using position and speed data provided by the vehicle's GPS and timetable information. More precisely: transfers are only made at those parts of the transport network where wireless coverage is available and where the vehicle is scheduled to remain stationary for a sufficient period of time for data transfer to be completed. With the information from the vehicle tracking system, the AVLM knows when it is at a WiFi point, with the timetable information the AVLM knows how long it should remain at that point, and with the vehicle's speed data provided by GPS, the AVML can initiate data transfer when the vehicle has stopped at that point, thus avoiding transmission errors.

5.2. The SQCM

Based on specification of the part of the transport network to be analysed—i.e., a stop, a route, a set of line services, etc.—and the temporal aspects of the analysis, this module obtains the data and parameters required for quality control. The SQCM, as shown in Figure 3, is executed in four steps: the first step, tracking, consists of obtaining the initial set of position data that will be used for quality control; the second step, filtering, ensures that quality control will be carried out using an integral data set, selecting reliable position data to that end; the aim of the third step is to obtain the passing times for each of the stops on the line services analysed; and finally, the fourth step consists of calculating the metrics used for quality control.

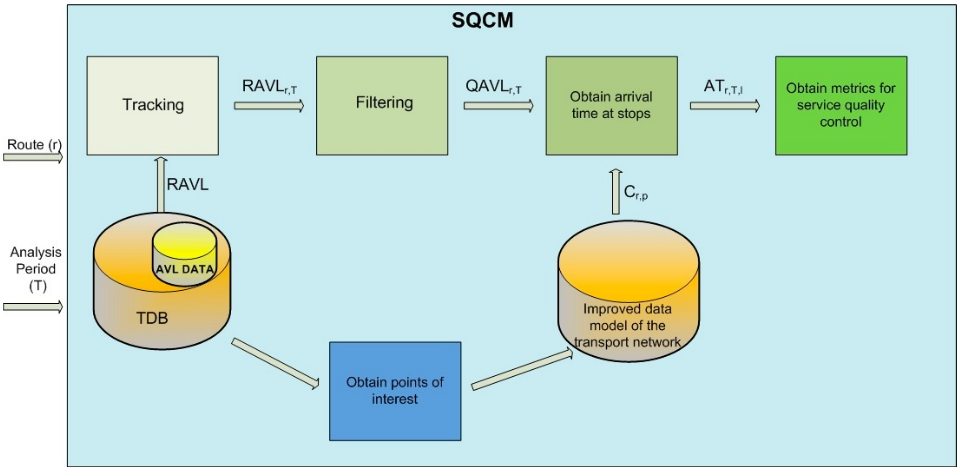

SQCM
Tracking
RAVL$_{r,T}$
Filtering
QAVL$_{r,T}$
Obtain arrival time at stops
AT$_{r,T,l}$
Obtain metrics for service quality control
Route (r)
RAVL
C$_{r,p}$
Analysis Period (T)
AVL DATA
TDB
Improved data model of the transport network
Obtain points of interest

Figure 3. SQCM Phases.

5.2.1. The Improved Transport Network Model

To calculate the various metrics used to evaluate the punctuality or regularity of a line service, the system must obtain the arrival time at each of the stops along the route. If the vehicles were only to stop at the bus stops, obtaining the arrival time at each stop would be simple. But in reality, the vehicles do not only stop at bus stops on the route, but also at unscheduled points due to traffic signs and congestion, specific incidents, etc. In order to distinguish when the position data for a stationary vehicle is due to the vehicle being at a scheduled stop or it having stopped due to traffic signs or congestion, the system must consider several scenarios. The least complex situation arises when a stop is at a distance greater than the maximum error in a GPS reading from the previous and next stops on the route, and also, the vehicle is not stopped due to a road sign or traffic conditions at a point on a radius greater than the maximum error of a GPS reading. The most complex situation arises when there is a stop and, next to it, at a distance less than the maximum error established by the system for GPS data, which was set at 25 m in the previous section, there is a traffic signal or a road feature—e.g., narrowing, roundabout or pedestrian crossing—which makes the vehicle stop. In this situation, the system must distinguish between GPS data for a stationary vehicle associated with it being at a bus stop, and the GPS data obtained when the vehicle was stationary due to traffic conditions or signs. In addition, it should be noted that because traffic conditions are variable, the points at which a vehicle might stop due to this factor are also variable. Figure 4 schematically represents cases of stationary vehicle data on a line service. The blue dots represent data for a vehicle when stationary at a stop. The red dots represent data for a vehicle that has stopped because of a traffic signal or congestion. The red circles represent the perimeter of radius U in which the vehicle is considered to be at a given point of the route. It is assumed that there can never be two stops at a distance below the threshold U. When a stop is near a traffic signal, there are areas in which stationary vehicle data are mixed owing to the vehicle being at a stop and because of the traffic signal.

In order to discriminate between the position data for stationary vehicles, our system uses an improved transport network representation model. The improvement consists in incorporating the points where vehicles stop systematically without being scheduled route stops. The technique used is based on a previous study published by the authors [32]. This technique uses the *k*-means classification method to classify all GPS stationary data obtained on the route under analysis and during the time interval in which quality control is being conducted. If the route has n stops, this classification process classifies all stationary data in m subsets, where $n \leq m$. Of these m subsets, there will be n subsets of GPS stationary data associated with scheduled stops and the remaining subsets will be GPS stationary data associated with unscheduled stops. This classification process is described below. The rules used to identify the clusters are as follows:

- Cluster associated with a scheduled stop on the route. If the cluster has a centroid very close to a point in the GPS data set and the scatter of the readings belonging to the cluster is low, then it is a set of readings associated with a scheduled stop of the vehicle. Because each point represents an area reserved for the vehicle to stop, a centroid is considered to be very close to a point if the distance between them is not more than 25 m (maximum error). The first two sets of points in Figure 4 are cases of this type on the route being analyzed.
- Cluster associated with a traffic signal on the route. If the centroid of the cluster is not close to any bus stops and the scatter of the readings belonging to the cluster is low, then it is a set of readings associated with a singular point in the route, with the centroid being the position that represents the traffic signal being identified. The third set in Figure 4 is a case of this type along the sample route.
- Unidentified cluster. If the cluster exhibits high scatter in the GPS readings, then it is a cluster with readings associated with more than one point at which the vehicle stops systematically (bus stop or traffic signal). In Figure 4, the fourth and fifth sets illustrate a case of this type, just like the last three sets. For each cluster of this type, a new iteration is performed, and each time, the K-means algorithm restricted to the points assigned to the cluster being studied is applied, taking as an initial approximation in the second and following iterations the points associated with bus stops that are known, that belong to the cluster and the centroid of the cluster obtained. Thus, the points will be grouped on the new centroids, each of which is located around the bus stops and traffic signal points of the cluster. The operation must be reiterated until a result is obtained for which the points are grouped at short distances from the centroids of their cluster. If the centroid of one of these clusters is very close (less than 25 m) to a bus stop point known, then the cluster represents readings of bus stop. In contrast, if the centroid is not close to a bus stops point known then the cluster is associated with a point of traffic signal that has been identified.

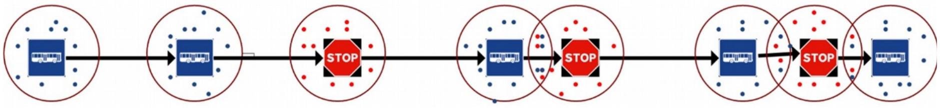


Figure 4. Cases of overlapping position data for stationary vehicles.

The route to be analysed for quality of service is r and P_r is the ordered set of n stops on the line. $RAVL_{r,T}$ is the set of GPS data obtained by all the vehicles that have completed route r during T, the period of time to be analysed. The subset $ZAVL_{r,T}$ is defined as the subset of $RAVL_{r,T}$, consisting of only GPS stationary vehicle data. Thus:

$$\forall RAVL \in ZAVL_{r,T} : RAVL \in RAVL_{r,T} \wedge RAVL_{vel} = 0$$

Once the $ZAVL_{r,T}$ set has been obtained, the first iteration of the k-means classifier is executed to split the $ZAVL_{r,T}$ set into n classes, taking as an initial approximation n centroids associated with each stop on the route. As a result n subsets of $ZAVL_{r,T}$ will be obtained, in which each subset i corresponds to the data belonging to class i, where $1 \leq i \leq n$. Each subset i is represented by $ZAVL_{r,T,i}$ and its centroid by $C_{r,T,i}$. If all $ZAVL_{r,T,i}$ clusters have only one stop and all GPS stationary data are at a distance from the centroid $C_{r,T,i}$ that is less than the threshold U, then the classification has converged in a solution in which we have reliably associated all stationary speed data with stops. Otherwise, if there is a cluster in which there is no stop, then that cluster contains GPS stationary data associated with unscheduled stops made by vehicles that have completed route r and there is at least one class with more than one stop. For each cluster with more than one stop the k-means classifier is only executed with the readings belonging to the class, and so on. Once all stationary data of a $ZAVL_{r,T}$ set have been classified, m classes are obtained, where $m \geq n$. The number of m classes is greater than n when at least one class has been obtained with GPS stationary data to which no stop belongs, and these readings are due to the vehicle stopping systematically at a point which is not a bus stop; we shall call these points unscheduled stop points on route r. The centroids of this type of cluster are added to the geographic database of the transport network as points of interest, to be taken into account when calculating the journey times of the different routes that pass through these points. The set of m centroids associated with route r is represented by $C_{r,T}$, and $C_{r,T,p}$ is a subset of this set and consists only of the centroids that represent the GPS stationary data associated with stops on the route. The subset $C_{r,t,p}$ will be used to obtain arrival times when vehicles stop at every stop p.

Below, Algorithm 1, we algorithmically describe the technique used to obtain the points of interest on route r, which as mentioned above are points where vehicles stop systematically without being stops on a route. The algorithm is recursive and uses three input parameters. The first parameter is a set of GPS coordinates for the route to be analysed, thus in the first invocation the $AVL_{r,T}$ set is used as the initial set of GPS data, consisting of the coordinates obtained by the vehicles when operating route r for time period T. The second parameter is the number of classes to be classified; in the first invocation the value of this parameter is the number of stops along the route. The third parameter corresponds to the initial approximations to each of the centroids of the classes; the input data in the first invocation are the GPS positions of the stops. As the output set, the $C_{r,T}$ set of centroids is obtained, representing each of the classes into which all the stationary data have been clustered.

Figure 5 illustrates a real-life result for this classification technique applied to the route studied in this article. It shows a map of part of the route with the position data, represented by blue and red dots, and the positions of two centroids of stationary data clusters, the red icons labelled with numbers 5 and 6, acquired during the analysed line services. Three stops on the route are identified by blue bus icons. The red dots represent GPS readings with zero horizontal velocity used by the classification technique to identify centroids of stationary data clusters.

We can also see from the map that there is a section between two of these stops containing a roundabout, hence the classification technique applied to stationary data on this section produces three clusters; following the route, these correspond to the following: the first, to the first stop of the section, the second, to the point where vehicles stop to join the roundabout—the red icon labelled with the number 5—and the third, to the second stop. Therefore, to calculate the arrival time at each of these stops, only the data pertaining to each of their clusters are considered, and data pertaining to the clusters represented by the centroid labelled with the number 5 are ignored. A similar case occurs with the stationary vehicle points in the proximity of the point represented by the red icon labelled with the number 6, which corresponds to a point where there is a stop sign and to the bus stop closest to this point. The technique that we have described classifies these points into two clusters, using only stationary vehicle data belonging to the cluster associated with the stop to obtain the arrival time at this stop.

Algorithm 1. Stationary data Classifier

```
Procedure Classifier_ Stationary_Data ($Q_r$ , $N$, $C_r$)
Input data:
    $Q_r$: set of position data obtained by vehicles operating line services on route $r$.
    $N$: number of classes to be classified.
    $C_r$: set of initial approximations to centroids.
Output data:
    $C_{r},_{T}$: centroids representing each of the classes of GPS stationary data.
Initial values:
    $Z_r$ = Ø;

Step 1: Obtain $Z_r$ which is the set of GPS stationary data on route $r$:
For each RAVL position reading belonging to $Q_r$,
    if $RAVL_{vel}$ = 0 then
        include $RAVL$ in $Z_r$.
    End if
Done

Step 2: Classifying by procedure K_means_classifier ($Z_r$, $n$, $P_r$, $Z_r$, $C_r$)
Input data:
    $Z_r$: set of stationary data to be classified.
    $n$: number of stops on the route.
    $P_r$: positions of the $n$ stops on route $r$.
Output data:
    $n$ $Z_r$ clusters
    $n$ $C_r$ centroids of each cluster

K_means_classifier ($Z_r$, $n$, $P_r$, $Z_r$, $C_r$ )

Step 3: Identifying the resulting clusters
For each $Z_{r,i}$ resulting cluster, where $1 \leq i \leq n$
    If $Z_{r,i}$ cluster contains no stop then
        Label the cluster as an unscheduled stop point.
    End if
    If $Z_{r,i}$ contains one and only one stop $p$ on the line then
        Select $Z_{r,i}$ as the set of readings to obtain the time of arrival at the stop on the line service of route $r$.
    End if
    If Zr,i contains k stops and k > 1 then:
        Classifier_ Stationary_Data ($Z_{r,i}$, $k$, $C_{r,i}$)
        Where:
            $Z_{r,i}$: the GPS stationary data belonging to class $i$.
            $k$: number of stops in cluster $i$.
            $C_{r,i}$: positions of the $k$ stops in $Z_{r,i}$.
    End if
Done
```

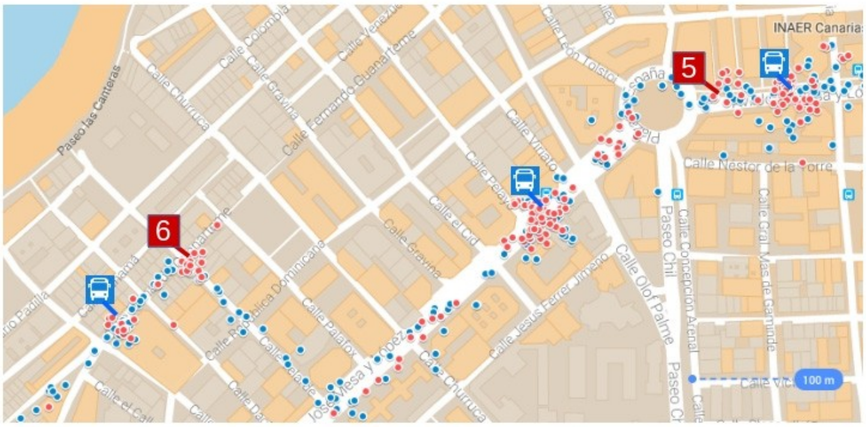


Figure 5. Example of a set of position data for vehicles in an urban section of the analysed route.

5.2.2. Step 1: Tracking

The quality control process starts with specification of the route, r, and the period of time, T, during which it is to be conducted. From these initial specifications, the objective of this step is to obtain the RAVL set of data acquired when the line services on route r were completed in time period T, a set hereinafter represented by $RAVL_{r,T}$. As the RAVL data are obtained periodically by all vehicles in the fleet while operating the line services, regardless of the service in question, $RAVL_{r,T}$ is a subset of the set that comprises all the RAVL data of the system. To select the RAVL data that are part of $RAVL_{r,T}$ it is necessary to first obtain the set of line services on route r that have been operated during the period in question, i.e., $L_{r,T}$. For each line service $l_i \in L_{r,T}$ we need to know the following: the vehicle that operated it and the time it began and ended. As we have already discussed in Section 3, all activity carried out by the transport operator is stored in the TDB. The data stored in this database include the start and end times of the line services, and this is where the data required for each $l_i \in L_{r,T}$ are obtained. This TDB query produces a data set in which each record represents a $l_i \in L_{r,T}$; Table 3 shows the structure of each of these records.

Table 3. Structure of the table of records obtained from the TDB query.

Field	Description
l_{veh}	Vehicle identifier
l_{T0}	Time that the line service on route r began, expressed in UTC.
l_{T1}	Time that the line service on route r ended, expressed in UTC.

- Once the required data from each $l_i \in L_{r,T}$ have been obtained, the $RAVL \in AVL_{r,T}$ data will be those for which the vehicle field, $RAVL_{Veh}$, coincides with the vehicle field of a $l \in L_{r,T}$ record and the time that the position was captured; $RAVL_T$ is greater than or equal to the time at which line service l, l_{T0} began, and less than the time at which line service l, l_{T1} ended. In formal terms:

$$\forall RAVL \in AVL_{r,T}, \exists! l \in L_{r,T} : RAVL_{Veh} = l_{Veh} \wedge l_{T0} \leq RAVL_T \leq l_{T1}$$

- For there to be a reliable relationship—based on time data—between the data from the TBD and the AVL module, both components must be synchronised using the same clock source. Therefore, the driver's console uses the GPS receiver clock to synchronise its clock.

5.2.3. Step 2: Filtering

For the quality control to be reliable, data integrity must be ensured; this is the purpose of this step in the process. Data integrity means that there are no erroneous data and that the dataset is complete. For our system, this integrity is achieved by executing filtering rules: the first, to eliminate from the $RAVL_{r,T}$ set those records containing position data that are not good quality; the second, to remove those records that are not part of a sequence of position readings that represent a complete and consistent route.

$RAVL_{r,T}$ records with low-quality GPS data are eliminated using the fields that indicate the age of the data, $RAVL_{Age}$, and type of data, $RAVL_{Sou}$; thus the RAVL record is deleted if either of the following two conditions are met:

- The data are old, i.e., $RAVL_{Age} = 1$, or no information is available for the age of the data, i.e., $RAVL_{Age} = 0$.
- Information on the type of data obtained (2D or 3D) is not available, i.e., $RAVL_{Age} = 0$.

$RAVL_{r,T}$ records that are not part of a complete and consistent route r are eliminated by checking that the sequence of RAVL records obtained for each line service $l_i \in L_{r,T}$ is complete, and that it represents a route consistent with route r, i.e., it passes through all the stops and in the planned order. In formal terms:

- The line service is $l \in L_{r,T}$, and $AVL_{r,T,l}$ is the set of position data obtained during line service l, where l_{T0} and l_{T1} are the start and end times of line service l, respectively, and Q the sampling period used to record the RAVL data during operation of the service. The data sequence $RAVL_{r,T,l}$ is then complete if the number of records in this sequence, $NAVL_{r,T,l}$ is:

$$NAVL_{r,T,l} = \frac{(l_{T1} - l_{T0})}{Q}$$

- $C_{r,T,p} = \{c_0, c_1, ..., c_n\}$ is the ordered set of centroids associated with each of the stops along route r. The route represented by $RAVL_{r,T,l}$ data is thus consistent with route r if:

$$\forall c_i \exists! RAVL : Dist\left(RAVL_p, c_i\right) \leq U,$$

where $Dist$ is the distance between two points, RAV_{Lp} the position logged in a $RAVL \in RAVL_{r,T,l}$ record, and U a distance threshold representing proximity. This value depends on the maximum GPS error accepted by the system, E, the length of the parking space at the bus stop, the maximum speed that the vehicle may attain and the time period used by the vehicles to acquire their GPS position data.

As a result of this step, an integral set of RAVL records is obtained, i.e., data containing position data of good quality and representative of line services on route r that are complete and consistent with route r as planned. Hereinafter, this set will be represented by $QAVL_{r,T}$. Similarly, the set of reliable data obtained during a line service l, on route r and operated during the period T will be represented by $QAVL_{r,T,l}$.

5.2.4. Step 3: Obtaining Passing Times

Once the $QAVL_{r,T}$ set has been obtained, the next step is to obtain the position data captured at the time that the vehicle arrives at each of the stops. To this end, the system partitions the $QAVL_{r,T}$ set into subsets of $QAVL_{r,T,l}$ data. If the position data for each of these $QAVL_{r,T,l}$ subsets are ordered on an ascending scale according to the time that each record was obtained, we would have a representation of a complete and consistent route r operated by the vehicle in question. The passing times for each of the stops on each of these $QAVL_{r,T,l}$ routes, which will be represented by the ordered time sequence $AT_{r,T,l}$, will be obtained considering two different cases:

- Case A: the vehicle passes stop p and does not stop. This case occurs when there are no passengers waiting to board the vehicle at p, and no passengers on the bus who want to alight at p. In this case, the passing time for this stop is obtained by finding the position for the $QAVL_{r,T,l}$ set that is at a minimum distance from the position of stop p. The distance threshold used in this case is the maximum error that the system assumes for a GPS reading, the variable E which was fixed at 25 m in Section 3.
- Case B: the vehicle stops at stop p. This case occurs when there are passengers at the stop or passengers on the bus who want to alight at p. In this case, the time of arrival at stop p is obtained by taking centroid $C_{r,p}$—obtained in the classification of GPS stationary vehicle data for route r as described in Section 5.2.1—as the stop position. The arrival time at stop p of a service line l on route r is the time at which the GPS position for the $QAVL_{r,T,l}$ set was recorded, and which must meet the following conditions:
    - ○ It is stationary.
    - ○ Of all the $QAVL_{r,T,l}$ stationary data, it is the nearest to $C_{r,p}$.

5.2.5. Step 4: Calculating the Metrics

Having obtaining the passing times for each of the stops on line service $AT_{r,T,l}$, the values of the various metrics used to evaluate the punctuality or regularity of the service line are then obtained. To monitor punctuality, for line services scheduled by timetable, the basic parameter is AD (1). From this value the remaining metrics used to evaluate punctuality may be obtained. To monitor regularity, for services scheduled by frequency, the basic parameter is HR (3). From this value the remaining metrics used to evaluate regularity may be obtained.

6. Use Case: Results and Discussion

Currently the system described in this article is in operation at the public transport company, Global Salcai-Utinsa S.A. This company is based on the island of Gran Canaria (Canary Islands, Spain), has a fleet of 345 vehicles, and operates 127 different routes on a transport network that contains 2686 stops. In 2015, this company carried out an average of 2395 line services on all routes, covering 28,897,002 km and carrying 19,284,378 passengers in that year. The system has been integrated with the other components used by this company on its onboard systems and its communications systems. The quality control results obtained for one of its routes are described below. The route itself is 23 km long, starts in the city of Las Palmas and ends in the city of Arucas; it has 30 stops, and it begins in an urban area in the city of Las Palmas, until the sixth stop, after which it becomes an intercity route, until the last four stops, where it again becomes an urban route through the town of Arucas. The company uses the code 210 to identify the route; therefore in the formalisation developed for this article, $r = 210$. Figure 6 gives an aerial view of the route with the stops represented by bus icons.

The results were obtained during period T—the whole of 2015—therefore in our formalisation it will be indicated as follows: $T = 2015$. Route 210 was scheduled by timetable during the analysed period. The AVL system recorded 51,499,404 position readings during all the line services completed for this route. The number of vehicles involved in these line services for the seleted route was 166, being 16 the number of vehicles that carried out these services systematically. In the preliminary analysis of the route to detect systematic stopping points that are not stops on the route, 7 such points were identified. Figure 7a displays these points as numbered red icons. These points are the centroids of a set of stationary vehicle data. Figure 7b,c are photographs of two of these points. Figure 7b corresponds to point number 5; this is a systematic stop point because the road on which the vehicle is travelling is a bus lane ending in a roundabout, and entry into the roundabout is controlled by a traffic light. Figure 7c is a photograph of point number 7; this is a point where the vehicle stops because it is on a slip road (right lane) that joins a dual carriageway (left lanes) and does not have right-of-way.

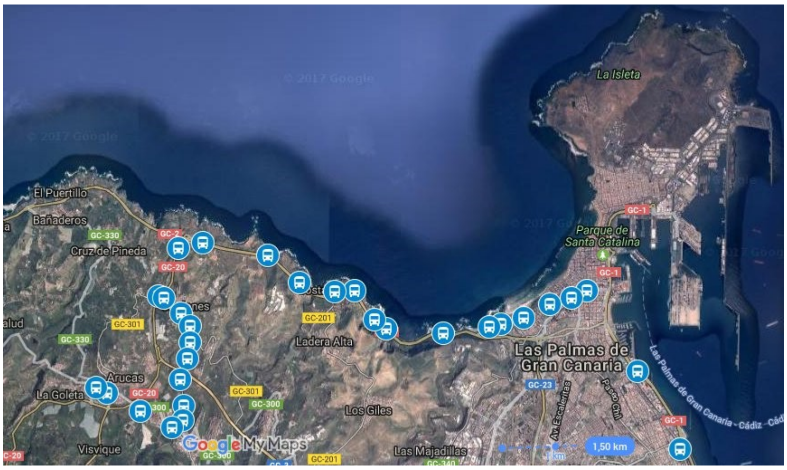


Figure 6. Aerial view of the route used for the case study. The bus icons represent stops on the route.

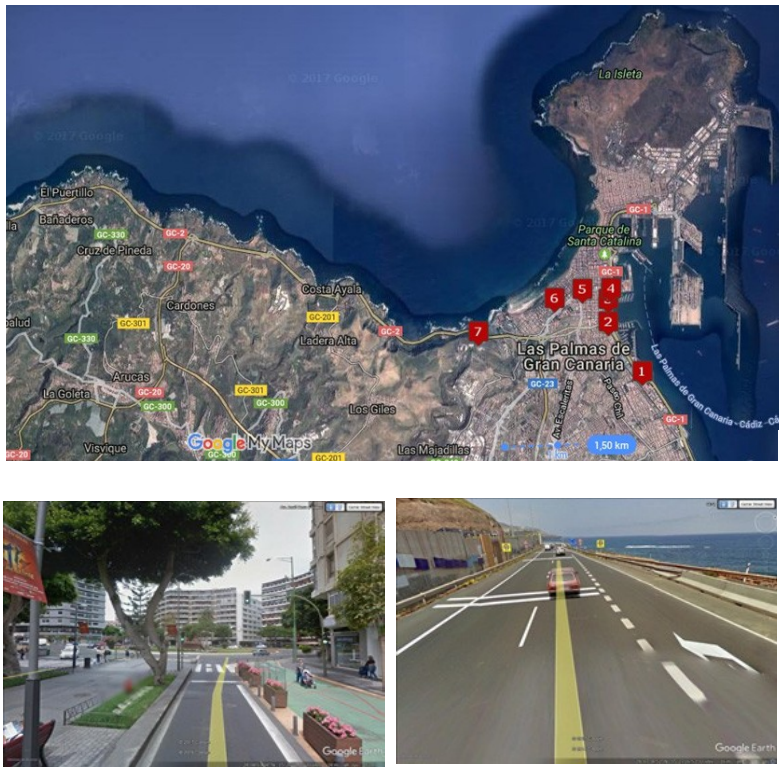


Figure 7. (**a**): aerial photo showing the location of the 7 unique points on the route. (**b**): photograph of point 5. (**c**): photograph of point 7.

The tracking phase was applied to this data set, and 617,195 position records were selected for 9675 line services on route 210 in 2015; these data form the set $AVL_{210,2015}$. The filtering phase was applied to this set of GPS data, and verified 8572 line services for which the recorded data sequences represented complete and consistent routes; hence the set $QAVL_{210,2015}$ contained 222,382 position records.

Once the GPS data had been filtered, we obtained the arrival times at each stop on each line service. The timetables of the line services on this route are scheduled by passing time, with two day types: one the one hand, weekdays (Monday to Friday), and on the other hand, weekends and public holidays. The results correspond to the two types: first, the results for weekday line services starting at 07:40 a.m., and then the earliest Sunday service, which starts at 8:40 a.m. Figure 8 shows the arrival times at each stop for the first set of line services, those that run from Monday to Friday. The horizontal axis represents the stops in order, i.e., the value 1 is associated with the first stop on the route, 2 with the second and so on until the last stop of the route, which has the value 30. The blue line represents the scheduled passing time, the average passing time is represented by the orange line, the earliest passing time by the yellow line, and the latest passing time by the grey line.

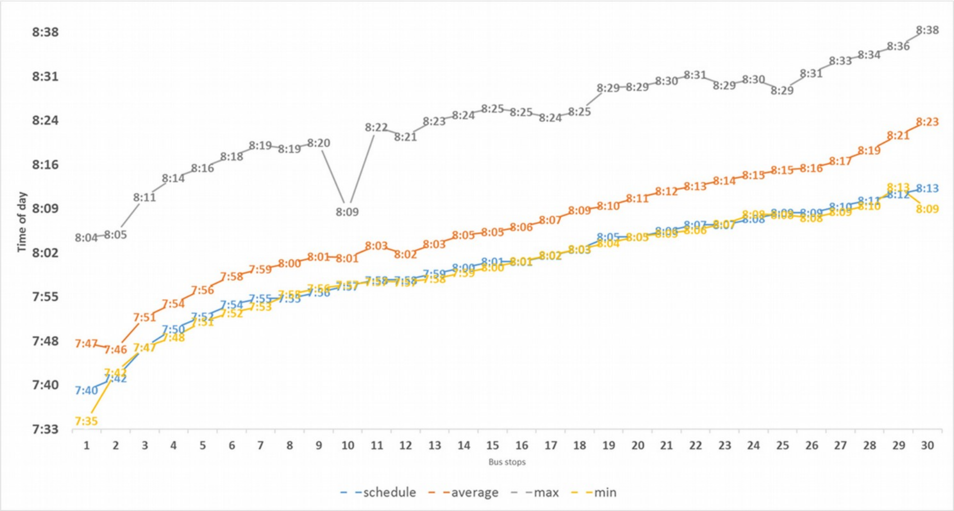
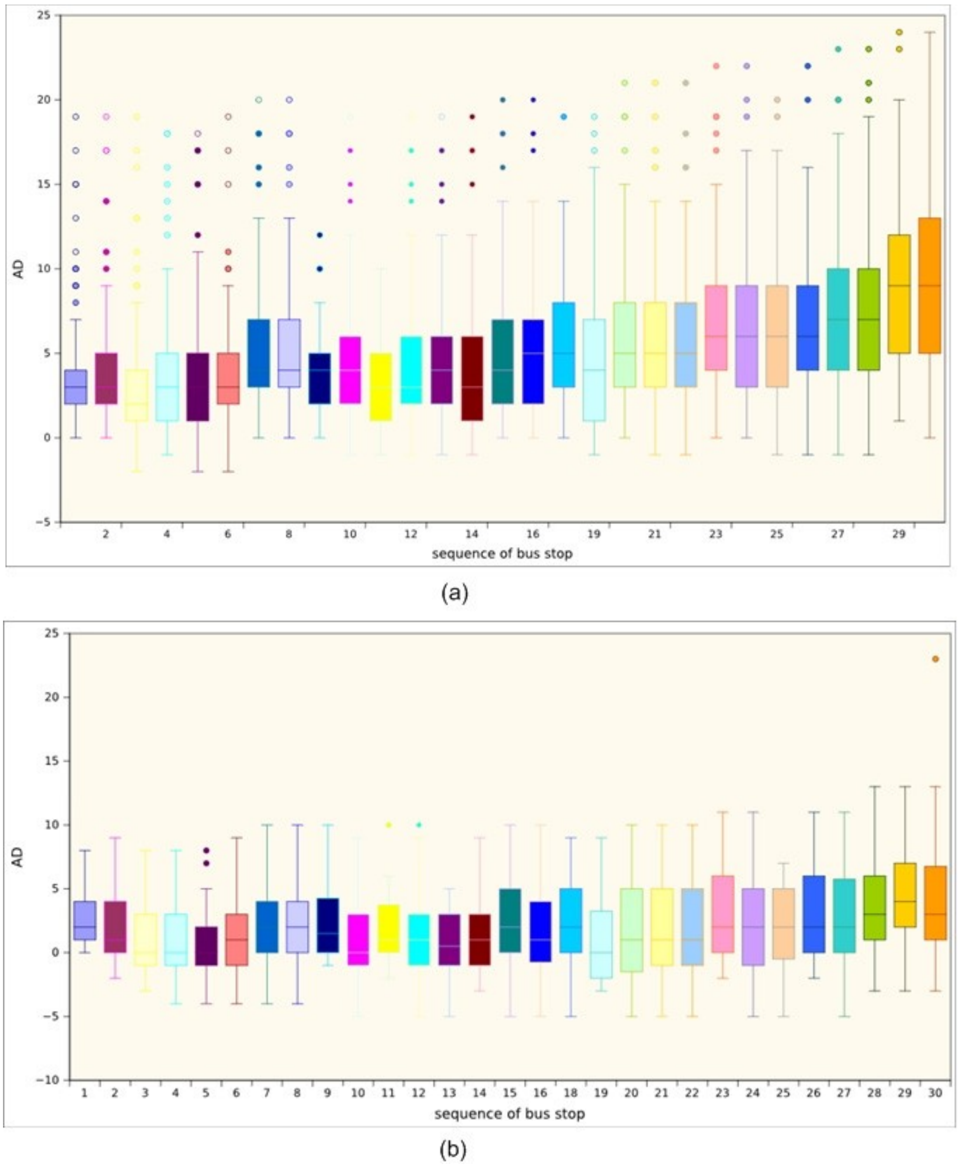


Figure 8. Passing times of line services starting at 07:40 Monday to Friday, excluding public holidays.

Figure 9 shows two box-and-whisker plots displaying the percentile distribution of the AD metric at each stop on the route for line services that start at 07:40 from Monday to Friday (plot (a)) and for each stop on all line services that start at 08:40 on Sunday (plot (b)). In the plots, the value 0 on the vertical axis represents perfect timetable adherence, each band inside the boxes represents the median value, i.e., the delay value below which 50% of the AD values are located. The circles represent outliers: very small, very large or very rare values.

As can be seen, the line services on Sundays are much more punctual than those that run on weekdays, possibly due to the difference in traffic congestion between weekdays and Sundays. We may also note that, for weekday line services and for most of the stops, delays continue or increase the further along the route the bus is, ranging from 2 to 8 min. By contrast, on Sunday line services, the increase and decrease in delays at stops does not follow a fixed pattern, and the median value is maintained between 0 and 5 min.The AD parameter values for each stop of the line service on the route are displayed in Figure 10, irrespective of day type and time. The median values for delay times are represented by colours: the green stops indicate a median value of less than 1 min, blue stops a median value between 2 and 3 min, light orange stops between 4 and 5 min, dark orange stops between 6 and 7 min and, finally, the red stops represent a median delay value of more than seven minutes.



(a)

(b)

Figure 9. Box-and-whisker plots showing median values for the AD metric at each of the stops on the route for line services that run from 07:40 from Monday to Friday (plot (**a**)) and from 08:40 on Sunday (plot (**b**)).

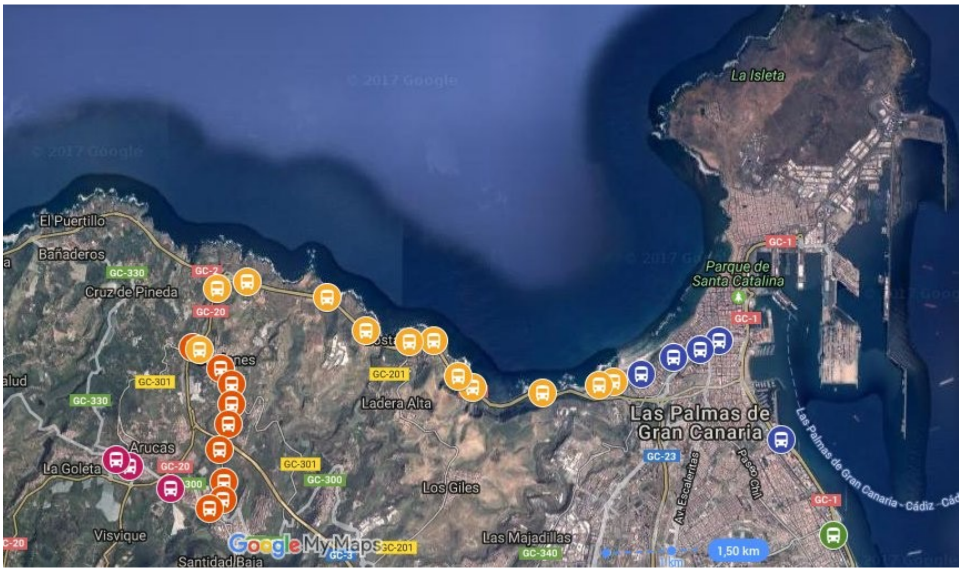


Figure 10. Stop categories according to delay, irrespective of day and time.

Figure 11 shows the RTV value for each type of line service. Graph (a) shows the values for weekday services (Monday to Friday). Graph (b) shows the values obtained for Sundays. Comparing the figures, it may be seen that punctuality varies according to the day type: the line service is less punctual on weekdays than it is on Sundays. The explanation for this is that there is less traffic on Sunday and there is lower demand at that time, so the buses run more smoothly and quickly, and are thus more punctual. We may also note that during the months of July and August, the weekday services are more punctual than during the rest of the year, possibly because those months are typically during the main holiday period, so there is less traffic and lower demand.

The system described in this article provides two main contributions. The first is that it is a complete, detailed and realistic description of a system for measuring the regularity and punctuality of regular public passenger transport by road, solving the challenge of massive data management involved in continuously monitoring regularity and punctuality. The description is complete and detailed because it includes all the formal and technological components and details of the system. It has been implemented with elements that are commonly used by transport operators, thereby facilitating implementation and deployment. The second contribution is the technique used to obtain the arrival times at the stops of the transport network. This technique is the *k*-means classification method, used to distinguish stationary data due to scheduled stops from stationary data due to the existence of an unscheduled stop, i.e., due to a traffic sign or traffic congestion or the condition of the road on which the vehicle is travelling.

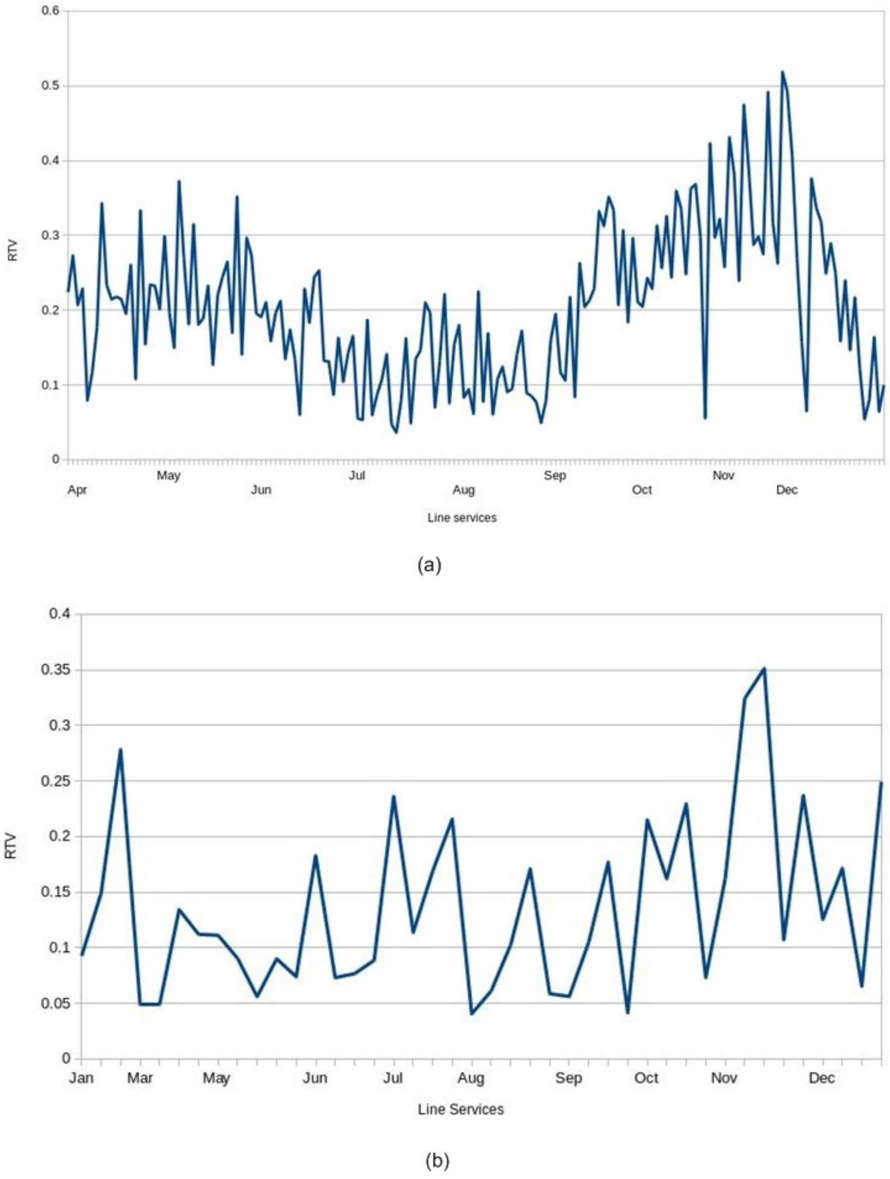


Figure 11. RTV metric values obtained during 2015. Graph (**a**) displays the values for each weekday service line that runs from 07:40. Graph (**b**) displays the values for each service line that runs on Sunday from 08:40.

Acknowledgments: The authors wish to express their gratitude to the Autoridad Única del Transporte de Gran Canaria (Transportation Agency of Gran Canaria) and to the public transport company Global Salcai-Utinsa S.A. They also declare that they have not received any funds to cover the publication costs of this article.

Author Contributions: C.R.G. directed the research study; all the authors contributed to the design of the system architecture; A.Q.-A. and F.A. developed the AVL system; G.P. and T.C. were responsible for developing the system data model and the service quality control module; all the authors contributed to system testing; all the authors participated in the preparation of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, J.; Wang, F.; Wang, K.; Lin, W.; Xu, X.; Chen, C. Data-driven intelligent transportation systems: A survey. *IEEE Trans. Intell. Transp. Syst.* **2011**, *4*, 1624–1639. [CrossRef]
2. Peters, A.; von Klot, S.; Heier, M.; Trentinaglia, I.; Hörmann, A.; Wichmann, H.E.; Löwel, H. Exposure to traffic and the onset of myocardial infarction. *N. Engl. J. Med.* **2004**, *17*, 1721–1730. [CrossRef] [PubMed]
3. World Health Organization. WHO Releases Country Estimates on Air Pollution Exposure and Health Impact. Available online: http://www.who.int/mediacentre/news/releases/2016/air-pollution-estimates/en/ (accessed on 24 January 2017).
4. European Commission. Road Safety: EU Reports Lowest Ever Number of Road Deaths and Takes First Step Towards an Injuries Strategy. Available online: http://europa.eu/rapid/press-release_IP-13-236_en.htm (accessed on 24 January 2017).
5. CEN/TC 320. *Transportation—Logistics and Services. European Standard EN 15140: Public Passenger Transport—Basic Requirements and Recommendation for Systems that Measure Delivered Service Quality*; European Committee for Standardization: Brussels, Belgium, 2006.
6. Peek, G.; van Hagen, M. Creating synergy in and around stations: Three strategies for adding value. *J. Transp. Res. Board* **2002**, *1793*, 1–6. [CrossRef]
7. Van Oort, N. Service Reliability and Urban Public Transport Design, Ph.D. Thesis, TRAIL Research School, Delft, The Netherlands, 2011.
8. Moreira-Matias, L.; Mendes-Moreira, J.; de Sousa, J.F.; Gama, J. Improving Mass Transit Operations by Using AVL-Based Systems: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 1636–1653. [CrossRef]
9. Yu, B.; Lam, W.H.K.; Tam, M.L. Bus arrival time prediction at bus stop with multiple routes. *Transp. Res. Part C Emerg. Technol.* **2011**, *19*, 1157–1170. [CrossRef]
10. Chang, H.; Park, D.; Lee, S.; Lee, H.; Baek, S. Dynamic multi-interval bus travel time prediction using bus transit data. *Transportmetrica* **2010**, *6*, 19–38. [CrossRef]
11. Jeong, R.; Rilett, L.R. Bus Arrival Time Prediction Using Artificial Neural Network Model. In Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems, Washington, DC, USA, 3–6 October 2004; pp. 988–993.
12. Turnquist, A. Strategies for improving bus transit service reliability. *Transp. Res. Rec.* **1982**, *818*, 7–13.
13. Polus, A. A study of travel time and reliability on arterial routes. *Transportation* **1979**, *8*, 141–151. [CrossRef]
14. Nakanishi, Y. Bus performance indicators: On-time performance and service regularity. *J. Transp. Res. Board* **1997**, *1571*, 1–13. [CrossRef]
15. Strathman, J.; Kimpel, T.; Dueker, K. Automated bus dispatching, operations control and service reliability. *J. Transp. Res. Board* **1999**, *1666*, 28–36. [CrossRef]
16. Barabino, B.; Di Francesco, M.; Mozzoni, S. Regularity diagnosis by automatic vehicle location raw data. *Public Transp.* **2013**, *4*, 187–208. [CrossRef]
17. Lin, N.; Yang, X.; Zhou, X.; Xu, X. The calculation of the punctuality rate between bus sites based on AVL data. In Proceedings of the 2013 Fifth International Conference on Measuring Technology and Mechatronics Automation, Hong Kong, China, 16–17 January 2013; pp. 1150–1152.
18. Furth, P.G.; Muller, T.H.J.; Strathman, J.G.; Hemily, B. *Uses of Archived Avl-Apc Data to Improve Transportation Management and Performance: Review and Potential*; Report 113; Transportation Research Board: Washington, DC, USA, 2003; pp. 1–12.
19. Riter, S.; McCoy, J. Automatic Vehicle Location—An overview. *IEEE Trans. Veh. Technol.* **1997**, *26*, 7–11. [CrossRef]
20. Theiss, A.; Yen, D.C.; Cheng-Yuang, K. Global positioning systems: An analysis of applications, current development and future implementations. *Comput. Stand. Interfaces* **2005**, *27*, 89–100. [CrossRef]

21. Zhou, H.; Hou, K.M.; Zuo, D.; Li, J. Intelligent urban public transportation for accessibility dedicated to people with disabilities. *Sensors* **2012**, *12*, 10679–10692. [CrossRef] [PubMed]
22. Mazloumi, E.; Currie, G.; Rose, G. Using GPS Data to Gain Insight into Public Transport Travel Time Variability. *J. Transp. Eng.* **2010**, *136*, 623–631. [CrossRef]
23. Zhao, J.; Dessouky, M.; Bukkapatnam, S. Optimal slack time for schedule-based transit operations. *Transp. Sci.* **2006**, *40*, 529–539. [CrossRef]
24. Derevitskiy, I.; Voloshin, D.; Mednikov, L.; Karbovskii, V. Traffic estimation on full graph of transport network using GPS data of bus movements. In Proceedings of the 5th International Young Scientist Conference on Computational Science, Krakow, Poland, 26–28 October 2016.
25. Cortes, C.E.; Gibson, J.; Gschwender, A.; Munizaga, M.; Zúñiga, M. Commercial bus speed diagnosis based on GPS-monitored data (2011). *Transp. Res. Part C Emerg. Technol.* **2011**, *19*, 695–707. [CrossRef]
26. Garcia-Castro, A.; Monzon, A. Using Floating Car Data to Analyse the Effects of ITS Measures and Eco-Driving. *Sensors* **2014**, *14*, 21358–21374. [CrossRef] [PubMed]
27. Dabove, P.; Manzino, A.M. GPS & GLONASS Mass-Market Receivers: Positioning Performances and Peculiarities. *Sensors* **2014**, *14*, 22159–22179. [CrossRef] [PubMed]
28. Hurn, J. *GPS: A Guide to the Next Utility*; Trimble Navigation: Sunnyvale, CA, USA, 1989; p. 46.
29. An Analysis of Global Positioning System (GPS) Standard Positioning System (SPS). Performance for 2013. Available online: http://www.gps.gov/systems/gps/performance/2013-GPS-SPS-performance-analysis.pdf (accessed on 24 January 2017).
30. Sinnott, R.W. Virtues of the Haversine. *Sky Telesc.* **1984**, *68*, 158.
31. Transmodel. Available online: http://transmodel-cen.eu/ (accessed on 24 January 2017).
32. Padrón, G.; García, C.R.; Quesada-Arencibia, A.; Alayón, F.; Pérez, R. Using Massive Vehicle Positioning Data to Improve Control and Planning of Public Road Transport. *Sensors* **2014**, *14*, 7342–7358. [CrossRef] [PubMed]

2.3. Systematic Approach to Analyze Travel Time in Road-based Mass Transit Systems Based on Data Mining

Received February 26, 2018, accepted May 3, 2018, date of publication May 17, 2018, date of current version July 6, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2837498

Systematic Approach to Analyze Travel Time in Road-Based Mass Transit Systems Based on Data Mining

TERESA CRISTÓBAL, GABINO PADRÓN, ALEXIS QUESADA-ARENCIBIA, FRANCISCO ALAYÓN, AND CARMELO R. GARCÍA
Institute for Cybernetics, University of Las Palmas de Gran Canaria, 35017 Las Palmas, Spain

Corresponding author: Carmelo R. García (ruben.garcia@ulpgc.es)

This work was supported in part by the Spain Ministry of Economy and Competitiveness and in part by the State Program for Research, Development and Innovation Oriented to the Challenges of Society.

ABSTRACT Road-based mass transit systems are an effective means to combat the negative impact of transport that is based on private vehicles. Providing quality of service in this type of transit system is a priority for transport authorities. In these systems, travel time (TT) is a basic factor in quality of service. This paper presents a methodology, based on data mining, for analyzing TT in a mass transit system that is planned by timetable. The objective of the methodology is to understand the behavior patterns of TTs on the different routes of the transport network, as well as the factors that influence these patterns. To achieve this objective, the methodology uses clustering techniques to process the GPS data provided by the vehicles of the public transport fleet. The results that were obtained when implementing this methodology in a public transport company are presented as a use case, demonstrating its validity.

INDEX TERMS Road-based mass transit systems, travel time, intelligent transportation systems, data mining, pattern clustering, global positioning system.

I. INTRODUCTION

According to the International Energy Agency, there were an estimated 900 million passenger light-duty vehicles on our roads worldwide in 2015, a figure that is projected to grow to 2 billion by 2040 [1]. There is widespread agreement that road transport systems based on the use of private vehicles have a negative impact. This impact includes degradation of the environment, health and safety on roads, aspects that are particularly pronounced in densely populated areas. The World Health Organization estimates that approximately 3 million people die every year due to health problems caused by pollution [2]. One way to mitigate the negative impacts associated with this type of transport system is to develop efficient public transport systems that provide quality of service. Intelligent Transport Systems (ITS) are an effective means to meet this challenge. Therefore, in modern societies and in the new paradigm of the smart city, ITS has a fundamental role to play.

For public road transport systems to be an alternative to transport based on the use of private vehicles, they must provide quality of service to make them attractive to the general public. In the context of road-based mass transit systems, one of the most important factors that affect quality of service is timetable adherence. Adherence means punctuality in the frequency between services or scheduled stop times. For it to be reliable, information is required on travel time (TT) behavior according to time and space parameters on the transport network. This paper presents a methodology based on data mining for analyzing TT behavior in a context of mass transit systems planned by timetable, based on the GPS data provided by the vehicles of the public transport fleet. The proposed methodology provides information on the TT behavior of the lines according to the time of year, time of day and section of the route. It also evaluates quality of service based on punctuality, according to criteria and metrics that are widely used by transport agencies and the academic community working in this field. In addition, it provides a useful framework for making TT forecasts. Thus, transport planning may be geared towards efficiency and quality of service and it becomes possible to provide reliable information to the public transport user. Specifically, the proposal consists of using classification techniques to study TT behavior in mass transit systems planned by timetable; the originality of this approach lies in the fact that existing works on this topic have mostly looked at planning by frequency. It should also be pointed out that the required data are commonly used by transport companies and agencies, and therefore do not require the deployment of infrastructure other than that which already exists in mass transit systems. The proposal is consistent with the current ITS paradigm: continuous observation of what happens in the transport network, continuous processing of the data produced by these observations and continuous improvement of the services provided, in order to make transport systems more efficient, safe, sustainable and adapted to the needs of users [3]. Moreover, the proposed methodology may be used to facilitate implementation of traffic control strategies that prioritize public transport vehicles [4], a key aspect of the smart city paradigm. The objective is to have better knowledge of travel time behavior that will enable subsequent studies to focus more on the routes and thus introduce measures to reduce this time or its variability.

The rest of this article is organized into five more sections. The second section lists works related to the proposed methodology. The methodology is described in the third section. Next, the results of a use case implementing the methodology in the study of the travel time of a bus line of a public transport company are presented. The fifth section is a discussion of the results, and the final section draws the conclusions.

TABLE 1. List of abbreviations.

DW	Dwell time
EXP	Journey on which GPS reading was taken
GMM	Gaussian mixture models
KNN	K-nearest neighbors regression
LAT	Latitude according to GPS reading
LIN	Line on which GPS reading was taken
LON	Longitude according to GPS reading
OT	Observed time of arrival at stop
QUA	GPS reading quality indicator
RT	Nonstop running time between two consecutive bus stops
RTD	Relative deviation in travel time for the section
RTV	Run time variation
SCAFC	Smart Card Automated Fare Collection system
ST	Scheduled arrival time
STO	Stop Node Identifier
SVM	Support vector machines
TD	Difference between the observed arrival time and the scheduled arrival time
TDB	Transport database
TIM	Time at which GPS reading was acquired, expressed in Coordinated Universal Time (UTC)
TT	Travel time
VEH	Vehicle
VEL	Vehicle speed at the GPS coordinates
VJ	Vehicle Journey
VJT	Date and time that VJ began, expressed in UTC
VLR	Vehicle location record

A. LIST OF ABBREVIATIONS

Table 1 contains a list of abbreviations used throughout this paper.

II. RELATED WORKS

In order to achieve the objectives of efficiency and quality of service in public transport, a fundamental requirement is to understand the mobility needs and habits of people. Based on this information, the three basic processes on which public transport is based—transport network design, service planning and operations control—may be carried out with guarantees. In [5] a global review was conducted of the methods used to design and schedule a transport network, and in relation to quality assessment, [6] provides an exhaustive review of the methods used to analyze behavior and to evaluate the main parameters that affect it. Technological advances, especially in mobile communications, sensors and computing, have enabled Intelligent Transport Systems to be developed that adapt transit systems to the needs of their users, to be more efficient and to provide greater quality of service. A common feature of this type of system is that it provides information on what happens in the transport network by performing an analysis of its time–space behavior from large amounts of data [7]. In this context of the massive use of data, data mining is a field that is increasingly used in transport engineering. A review of the literature in which data mining has been used to solve some of the problems in transport systems is presented below. Depending on the data source used, these works may be classified into two groups: those based on data related to the movements of travelers in the transport network and those that use data related to the location of the vehicles in the transport network. In both groups there are works that address the three main problems that need to be addressed to achieve efficiency and quality of service.

The works that use data associated with the movements of travelers include studies that: seek to acquire information about the profiles and usage habits of transport network users [8]; measure the use of the network infrastructure by travelers [9]; make predictions about travel times and develop personalized information services for the user [10], [11], based on records generated by the use of the Smart Card Automated Fare Collection system (SCAFC). In [12], socio-demographic factors are also taken into account: location of shopping centers, sports areas, residential areas, etc. The works that propose techniques to obtain mobility patterns of mass transit system users may be grouped into two categories according to the analysis carried out in [13]: those based on statistical methods capable of supplying a self-explanatory model when treating them as the result of a stochastic process, and those that use neural networks. In order to predict total demand in transit systems, based on time series of trips completed during certain time intervals, the use of statistical models is proposed in [14] and neural networks are used in [15]; as an example of mixed procedures, a process to select the generated functions before applying the neural network is introduced in [16]; in [17], the result of two different models of networks is analyzed using time-dependent parameters (trend, cycle and periodicity) in the observed demand data; and a new hybrid optimization algorithm is developed in [18], with set theory and neural network techniques, to predict the volume of passengers by road. As an example of other methods, in [19] the space–time behavior of travelers in a metro network based on the use of cards is studied using clustering techniques.

The location data of public transport vehicles have been used mainly to improve the design of the transport network, to evaluate the quality of service and to make travel time forecasts. The following works are examples of how different issues are tackled with these data: [20] proposes a methodology to evaluate the road network from the point of view of travel time stability through statistical distribution functions. In [21], by means of clustering techniques developed by the authors, the impact of demand and traffic on operational performance is analyzed. By gathering information on passengers boarding and alighting from vehicles, [22] looks at how to avoid overcrowding, which, together with delays in arrival times, can dissuade people from using public transport services; and in [23], diagnostic diagrams of service reliability are generated to determine how the variability of service attributes affects the behavior of travelers. In [24], a methodology for improving the design of the transport network is proposed: it detects the stop, classifies it, generates routes and estimates stop times by processing the vehicle GPS data using clustering techniques. Reference [25] proposes a new metric to evaluate the punctuality of buses using vehicle location data. In [26], the causes of scheduling irregularities are analyzed. In the context of road-based mass transit systems planned by frequency, in [27] and [28], location and passenger movement data are processed using the Gaussian Mixture Models (GMM) clustering technique and ad-hoc metrics with the aim of selecting the best cluster to evaluate quality of service taking into account the day coverage.

With regard to travel time forecasts by processing location data using machine learning techniques, a wide range of studies have been conducted on this subject. In [29], neural networks are used, and classification techniques are used in [30] with k-nearest neighbors regression (kNN), and in [31], k-means and v-means clustering. There are also a considerable number of proposals that tackle travel time predictions using state models and time series. For example, state models, more specifically Kalman filters, are used in [32] and time series in [33] and [34]. Lastly, a hybrid model using Support Vector Machines (SVM) and Kalman filters is proposed in [35].

III. METHODOLOGY

This paper was developed in the context of road-based intercity or long-distance mass transit systems. In this type of system, TT is an important criterion for providing quality of service. Firstly, because travelers want their journeys to last as little as possible, and secondly, because travelers expect punctuality. To achieve these objectives, accurate travel time estimates are needed, and the problem that arises when making these estimates is that travel time depends on factors such as traffic conditions, the travelers who board or alight from the vehicle at each stop on the route, the weather conditions at the time of the bus journey, etc.

This section describes a methodology for systematically analyzing travel time to improve quality of service and identify the factors that affect the travel time on each journey. Identifying these factors makes it possible to obtain TT behavior patterns for the different routes of the transport network and acquire a more precise knowledge of how this time varies, and thus:

- Improve the design of the transport network. Once the factors that affect the routes are known, the routes can be redesigned to reduce travel time.
- Plan more reliably. If the variations in travel time are known, more accurate estimates may be made.
- Control operations more efficiently. If the factors that affect travel time on a route and how this time varies depending on when the route is traveled are known, this greater understanding will enable real-time measures to be adopted that guarantee quality of service.
- Improve quality of service. If travel times are reduced and the reliability of service planning increased, quality of service will improve.

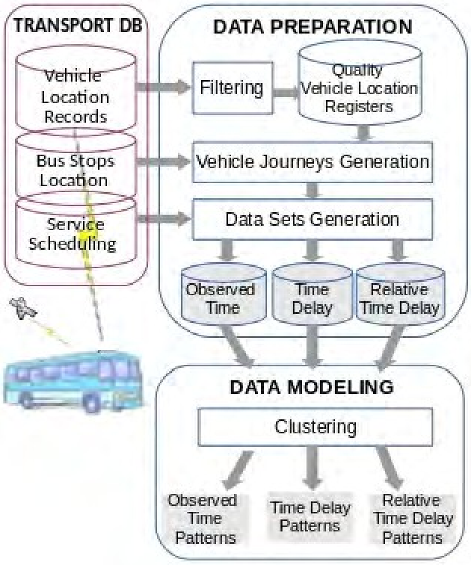


FIGURE 1. General diagram of the methodology.

The methodology used, as illustrated in Fig. 1, was inspired by the process-oriented methodology called CRossIndustry Standard Process for Data Mining (CRISP-DM) [36]. This study mainly followed two stages of this methodology: data preparation and modeling. In the first—data preparation—the data that form the basis of this study were merged and complemented. The second phase—modeling—incorporated modeling tools, based on clustering techniques, which were used to identify the factors that affect the travel time of a route and to obtain its patterns of behavior.

A. FORMALIZATION

This section describes the formal model used to analyze TT. This formalization had two objectives. The first was to ensure that the methodology can be applied to different types of road-based mass transit systems, and was achieved by applying standards related to conceptual models of public transport systems. The second was to ensure that the information provided is useful, and was achieved by using criteria and metrics for evaluating quality of service in public transport that are widely used by transport agencies and the academic community.

To achieve the first objective, the methodology was based on Transmodel (Reference Data Model for Public Transport) [37]. At the first level of formalization, the transport network was represented. At this level, the entities involved are the nodes and the arcs that physically link the nodes. The nodes represent places in the transport network where transport-related activity takes place: passengers boarding and alighting, schedule controls, ticket sales, etc. Each of these activities are attributes of the entity node, which is represented by n_i, with the subscript i being the identifier of the node. The set of nodes of the transport network is denoted by N, $N = \{n_i\}$. The N nodes are connected by arcs that represent the routes taken by vehicles and travelers, giving rise to a directed graph; this set of arcs is represented by A and gives rise to the physical graph of the transport network, represented by G, $G = (N, A)$. For the purposes of the methodology, the nodes of interest are the passenger boarding and alighting nodes, which will be given the generic name of *stop*, and the schedule control nodes. The set of nodes that fall within these categories is represented by P, where $P \subseteq N$. The arcs of interest are those that represent the routes followed by the buses carrying passengers; the set formed by arcs of this category is represented by W, where $W \subseteq A$. P. On this first level of formalization, the next entity is the route, which is defined as the path followed by the vehicles of the fleet, and comprises an ordered sequence of arcs. Each route is represented by r_i, where the subscript i is the identifier of the route. If route r_i has n arcs, then r_i is specified by n-tuple $(a_i, \ldots, a_n)$, where $a_i, \ldots, a_n \in W$. From the route entity, the line entity is defined: a set of very similar routes from the topological point of view (usually a round trip) represented by l_i, where the subscript i is the identifier of the line. The set of lines in the transport network is represented by L, $L = \{l_i\}$.

The second level of formalization is associated with service scheduling. This scheduling, represented by S, is organized into basic planning units, s_i, so $S = \{s_i\}$. Each s_i planning unit is defined as a set of ordered operations, in which the start and end times and the nodes at which the route starts and ends are specified. For the methodology, there are two aspects of interest at this level. The first is how to specify the calendar dates on which an s_i service must be performed, and the second is the minimum unit of time to be considered in the schedule plan. Specification of the calendar dates is done through a schema in which the different types of calendar day are described. Examples of the most common types are: day of the week, work day, public holiday, weekend, school period, etc. As for the smallest unit of time used for scheduling, this is usually a minute. Each of the completed operations of a line service—a route completed by the corresponding vehicle—is called a Vehicle Journey, represented by VJ. The *VJ* set of all the routes of a line l is represented by VJ_l; to express the *VJ* set of a line l completed in a period of time T the notation VJ_{lT} is used.

Of special interest are the criteria and metrics used to evaluate quality of service in mass transit systems where two types of schedule are distinguished: those based on frequency of service and those based on timetables. The first type is used in urban or short-distance transport [38]. The second type is used for intercity or long-distance transport, where TT and punctuality are two basic criteria for assessing quality of service. In general, the travel time of a VJ on route r, represented by *TTr*, is formally expressed as a function of two times: the dwell time at each stop, *DW*, and the time, *RT*, that it takes to cover each arc of the route:

$$TTr = \sum_{n=1}^{N_s} DW_n + \sum_{n=1}^{N_a} \mathrm{RT}_a \tag{1}$$

Where N_s is the number of stops on r_i, DW_n the time the vehicle remains stationary at stop n of the route (dwell time), N_a the number of arcs on the route and RT_a the travel time of arc a of the route.

The methodology was developed to analyze travel time in a context of an intercity or long-distance mass transit system. For this type of system, a metric used to evaluate punctuality is Run Time Variation (RTV) [39]. The calculation of this metric is expressed below:

$$\mathrm{RTV} = (N_p)^{-1} x \sum_{n=1}^{N_s} \frac{|OT_n - ST_n|}{OT_n} \tag{2}$$

Where N_s is the number of stops on route r_i, OT_n the observed time of arrival at stop n and ST_n the scheduled time of arrival at stop n. The value $(OT_n - ST_n)$ is the deviation from the scheduled arrival time at stop n, which is represented by TD_n. If this value is positive, it indicates a delay with respect to the planned time, a value of zero indicates that the arrival time at the stop is on schedule, and a negative value indicates the vehicle has arrived at the stop ahead of time.

Another aspect that the methodology takes into account is the cost incurred by non-adherence with VJ timetables [40]. In the case of timetabled bus lines, it is assumed that the traveler arrives at the stop moments before the vehicle is scheduled to pass by [41]. Therefore, the methodology assumes that the cost of non-adherence with a VJ timetable, represented by the variable *COST*, is the time cost for the travelers on that VJ of having to wait at the stop, represented by TI_n.

$$\mathrm{COST} = \sum_{n=1}^{N_p - 1} |OT_n - ST_n| \times TI_n \tag{3}$$

B. DATA PREPARATION PHASE

The objective of this phase is to obtain the basic data required to analyze travel time. These data are obtained from the records stored in the transport database (TDB). For the purposes of this methodology, the entities of interest in the TDB represent the activities that are planned and carried out in the transport network, comprising the following data:

- Geographical location of line stops. These locations are given by their GPS coordinates: latitude and longitude.
- Estimated stop arrival times for each planned VJ. This information is necessary to evaluate the punctuality during the period of analysis.
- The data that represent relevant events that occurred during the VJ, especially periodic updates of the vehicle GPS coordinates during the line services and the data used to ensure integrity.
- The total number of passengers that board and alight at each stop on the route, obtained from traveler payment records.

TABLE 2. VLR data structure.

VEH	TIM	LIN	EXP	LAT	LON	VEL	QUA

The basic data for this methodology are supplied by the readings that indicate the location of the vehicle at a given moment in time. The location of each vehicle is acquired periodically and stored in a data structure named Vehicle Location Record (VLR), this structure is shown in Table 2. The set of all location records is represented by $\{VLR\}$. The subset of $\{VLR\}$, comprising the locations obtained for vehicles on line l routes, in the period T, is represented as $\{VLR\}_{l,T}$. The set $\{QVLR\}$ is obtained from $\{VLR\}$. $\{QVLR\}$ is an integral dataset that guarantees the reliability of the results obtained by the methodology. In this case, integrity means that all records in the dataset $\{QVLR\}$ comply with the following properties:

- They contain a GPS reading of good quality, meaning that it was obtained using the signal provided by at least three GPS satellites and that the reading is less than 10 seconds old. These data properties were obtained from the protocol data used by the vehicle's GPS receiver.
- They were obtained on a VJ that completed a route, meaning that it went through all the planned stops in a coherent fashion (having covered all the planned arcs).

All the data for the $\{QVLR\}$ dataset were acquired through a filtering process (see Fig. 1). The subset of $\{QVLR\}$, comprising the locations obtained for vehicles on line l routes in the period T, is represented as $\{\mathrm{Q}VLR\}_{l,T}$. From the $\{QVLR\}$ dataset, the arrival times at each of the stops on the route are obtained for each VJ; $\{OT\}$ represents the arrival time dataset.

From all the arrival times for all the VJs, the three datasets—$\{OT\}_{l,T}$, $\{TD\}_{l,T}$ and $\{RTD\}_{l,T}$—to be used in the next phase of the data mining project, the modeling phase, were constructed for each line l at each instant of time T.

The dataset $\{OT\}_{l,T}$. This set was used to obtain the behavior patterns of arrival times of the line analyzed during the time period T. Table 3 shows the structure of dataset $\{OT\}_{l,T}$.

TABLE 3. Structure of the data associated with each element of the dataset *{OT}*.

VJ	VEH	VJT	STO	OT

The dataset $\{TD\}_{l,T}$ was obtained by calculating, for each data record, the deviation of the recorded arrival time from the scheduled arrival time. This dataset was used to obtain the behavior pattern of the deviations of the arrival times from the schedule and is therefore an indicator of the cost of said deviations. Table 4 shows the structure of dataset $\{TD\}_{l,T}$.

TABLE 4. Structure of the data associated with each element of the dataset *{TD}*.

VJ	VEH	VJT	STO	DT

The dataset $\{RTD\}_{l,T}$ (Relative Time Delay), was obtained by applying the following transformation to each data record of $\{TD\}_{l,T}$: Let TD_n be the deviation in the recorded arrival time of a vehicle on a VJ at stop n, and let $DTn - 1$ be the deviation at stop $n - 1$ on that same VJ, thus defining the relative deviation in the arrival time at stop n on the VJ (denoted as $RTDn$) as $RTD_n = TD_n - TD_{n-1}$. A positive value for RTD_n means that the vehicle has traveled the section between stops $n - 1$ and n in a time longer than planned, a value equal to zero means that the vehicle has traveled the section in the planned time and a negative value means that the vehicle has traveled the section in less time than planned. The dataset $\{RTD\}_{l,T}$ was used to identify the sections that cause the VJ to run early or late and to understand the pattern that the relative deviations follow in these sections. Table 5 shows the structure of dataset $\{RTD\}_{l,T}$, thus avoiding a cumulative effect in the TD.

TABLE 5. Structure of the data associated with each element of the dataset {R*TD*}.

VJ	VEH	DAT	STO	RDT

C. MODELING PHASE

The objective of this phase is to gain an understanding of the TT behavior of the VJ according to different variables. For a traveler on the route, TT is the time consumed in going from the origin stop to the destination stop of their journey. The ultimate goal is to understand the TT behavior at each and every stop on the route. Specifically, the aim is to understand how certain time-dependent factors, such as the type of calendar day and the time of day, affect the behavior of these times. Also, to understand how the deviations from the scheduled arrival times at stops, TD_n, develop depending on the section of the route. A final objective is to identify on which sections of the route deviations from the scheduled TT are generated according to the type of calendar day, and time of day.

The methodology employed clustering techniques to group the different TT data according to their similarity; this type of clustering technique was chosen because they are capable of handling large datasets that are frequently used in the context of transport, specifically the k-medoids algorithm [43], which is one of the most robust against noise. A medoid may be defined as the element of a group whose average dissimilarity to all elements in the group is minimal. It is the point located the closest to the center in the whole group.

Once a cluster solution has been obtained, the next step is to evaluate the validity of the solution. There are various ways of carrying out this evaluation, which may be broken down into three categories [44]. The first category consists of techniques based on external metrics, which measure the coincidence of the groups with previously generated labels for that class. The second category consists of techniques that use internal metrics, which measure the intrinsic information of each dataset. Finally, the third category consists of techniques that use relative metrics, which are based on the comparison of several different clustering solutions. For the purposes of this study, an internal index was chosen to measure the quality of the clusters in the first instance: the silhouette function [45]. This measures the consistency of the segments generated, based on the tightness and separation of its elements, and is computed by the following formula:

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & \textit{if } a(i) < b(i) \\ 0, & \textit{if } (a(i) = b(i)) \\ \dfrac{b(i)}{a(i)} - 1, & \textit{if } b(i) < a(i) \end{cases} \tag{4}$$

In Formula (4) $a(i)$ is the average distance from object i to the other objects within the cluster and $b(i)$ is the smallest average distance from i to all the objects of each of the clusters to which i does not belong.

The k-Medoids clustering technique was applied to datasets $\{OT\}_{l,T}$, $\{TD\}_{l,T}$ and $\{RTD\}_{l,T}$.

IV. RESULTS

This section presents the results obtained in a use case of the methodology. The use case consisted of analyzing the behavior over one year of the arrival times at the stops on a line of the public transport company Global Salcai-Utinsa. This is a company that operates on the island of Gran Canaria (Canary Islands, Spain) and is the main intercity transport company on this island; it has a fleet of 304 vehicles operating on a transport network with 2686 stops, 110 different routes and 2395 daily routes. Every year, its vehicles travel around 25,000,000 kilometers, transporting 20,000,000 passengers.

With regard to the tools used, in the data preparation phase, Oracle was used for the database system and Pentaho for integration and visualization. In the modeling phase, the RStudio framework was used; more specifically, the PAM function of the Cluster package [46], selecting the Euclidean distance as the metric for calculating the dissimilarities between the data and without determining the initial medoids.

TABLE 6. Scheduled VJs on the analyzed line service Monday to Friday (excl. public holidays).

VJ ID	Departure time	VJ ID	Departure time	VJ ID	Departure time	VJ ID	Departure time
1	06:30	9	10:40	17	14:40	25	18:40
2	07:10	10	11:10	18	15:10	26	19:10
3	07:40	11	11:40	19	15:40	27	19:40
4	08:10	12	12:10	20	16:10	28	20:10
5	08:40	13	12:40	21	16:40	29	20:40
6	09:10	14	13:10	22	17:10	30	21:30
7	09:40	15	13:40	23	17:40	31	22:15
8	10:10	16	14:10	24	18:10		

The line selected was number 210, and all the VJs of this line follow the same route. With regard to the route followed by the bus line, it should be noted that it starts in the city of Las Palmas de Gran Canaria, which is the island's main traffic hub, and ends in the city of Arucas, one of the largest population nuclei on the island. It crosses urban areas and non-urban areas, and some of its stops are near health, educational and commercial centers. Therefore, it is an illustrative use case of a bus line since the travel times of its different VJs are affected by different factors related to demand, the calendar, the time of day, traffic conditions, etc. Fig. 2(a) shows an aerial view of the line service with the location of each of its stops, with those considered significant for this study highlighted in red, as will be explained below. In Fig. 2(b) the same bus line is represented schematically, distinguishing between the different types of road that it transits. The route has 30 stops and one control point, and covers a distance of 23 kilometers. The period studied was the whole of 2015. In this period, the VJs were scheduled according to three types of calendar day: the first (type 0), was Monday to Friday, excluding public holidays, the second (type 1), Saturdays, and the third (type 2), Sundays and public holidays. Table 6 shows VJ schedule planning on the days pertaining to the first type (Monday to Friday, excluding public holidays). For this type of day, it may be seen that the first VJ started at 06:30, that between 07:10 and 20:40 a VJ started every 40 minutes, and that the last two VJs started at 21:30 and 22:15. Table 7 shows VJ schedule planning on the days pertaining to the second and third type (Saturdays, Sundays and public holidays). On these days, a VJ was scheduled for every hour between 08:40 and 20:40, with the last two VJs starting at 21:30 and 22:15.

TABLE 7. Scheduled VJs on the analyzed line service for Saturdays, Sundays and public holidays.

VJ ID	Departure time	VJ ID	Departure time	VJ ID	Departure time	VJ ID	Departure time
1	08:40	5	12:40	9	16:40	13	20:40
2	09:40	6	13:40	10	17:40	14	21:30
3	10:40	7	14:40	11	18:40	15	22:15
4	11:40	8	15:40	12	19:40		

As for the arrival time at each of the stops, the schedule provided for the same travel time for each stop regardless of the type of day and time of day of the VJ. The smallest

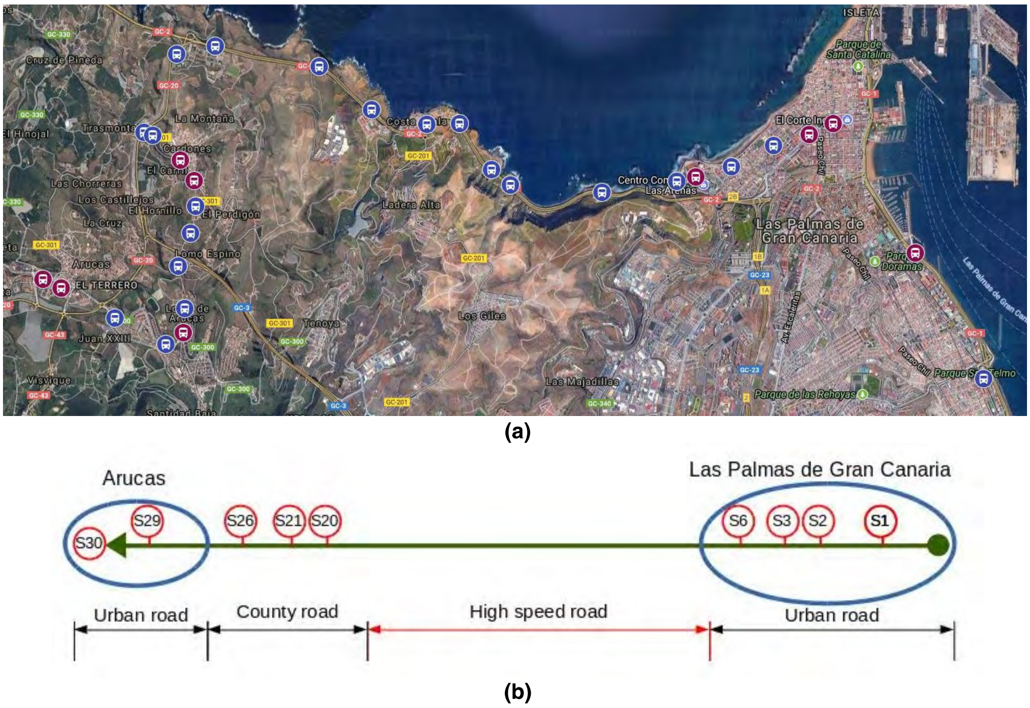


FIGURE 2. (a) Aerial view of the stops on line 210 (those considered significant in red). (b) Schema of the type of sections on the line.

TABLE 8. Arrival time at each stop on the line.

Stop	Arrival time	Stop	Arrival time	Stop	Arrival time
1	2	11	18	22	27
2	7	12	19	23	27
3	10	13	20	24	28
4	12	14	21	25	29
5	14	15	21	26	29
6	15	16	22	27	30
7	15	18	23	28	31
8	16	19	25	29	32
9	17	20	25	30	34
10	18	21	26		

TABLE 9. Number of users of the bus line that boarded or alighted at each stop in 2015.

Stop	Passengers	Stop	Passengers	Stop	Passengers
0	42960	10	13	21	8304
1	4856	11	13	22	1833
2	22772	12	2279	23	1530
3	9284	13	3472	24	704
4	3700	14	52	25	3301
5	2494	15	4446	26	10534
6	19571	16	1660	27	4047
7	251	18	2737	28	1496
8	14	19	210	29	9764
9	62	20	8301	30	42840

unit of time established in this schedule is a minute. Table 8 shows the planned arrival times for each of the 30 stops on the route. The first stop, stop 0, is not included in Table 8 since it is assumed that the vehicle starts the VJ at the scheduled time. The control point, labeled number 17, has also not been included in the table. Each stop on the line has been identified in the order of arrival following the set route; the stops correspond to the labeled points 0 to 16 and 18 to 30.

According to data from the TDB, the 10 stops that were most used by the passengers on this line in 2015 were: 0, 1, 2, 3, 6, 20, 21, 26, 29 and 30. Table 9 shows the number of passengers that board and alight at each of these stops. In Fig. 2(b) these stops are represented with numbered icons; the number indicates the order of that stop on the route, with the exception of the origin stop.

A. RESULTS OF THE DATA PREPARATION PHASE

Table 10 shows each dataset that was processed in the data preparation phase. These data refer to the year studied (2015). The total number of position readings obtained from the entire vehicle fleet after completion of all the VJs of all the lines defined in the transport network was 51,499,404.

TABLE 10. Number of elements of the datasets used in the methodology.

Set	Number of elements
$\{VLR\}_{2015}$: Total number of location readings obtained in 2015	51,499,404
$\{VJ\}_{210,2015}$: Number of VJs planned for line 210 in 2105	9675
$\{QVJ\}_{210,2015}$: Number of complete and coherent VJs completed on line 210 in 2015	6092
$\{QAVL\}_{210,2015}$: Number of location readings obtained on the VJs in dataset $\{QVJ\}_{210,2015}$	158,300
$\{OT\}_{210,2015}$: Observed arrival times processed in the modeling phase	6092
$\{TD\}_{210,2015}$: Deviations from the arrival times processed in the modeling phase	6092
$\{RTD\}_{210,2015}$: Relative deviations from the times processed in the modeling phase	6092

During this year, 9675 VJs were scheduled for the analyzed bus line. Of these 9675 scheduled VJs, when applying the filtering process, 6092 VJ were classified as complete and coherent from 158,300 location readings. From this integral set of location readings, the three datasets used in the modeling phase were created: $\{OT\}_{210,2015}$, $\{TD\}_{210,2015}$ and $\{RTD\}_{201,2015}$.

B. RESULTS OF THE MODELING PHASE

The objective of this phase is to obtain a pattern that describes the travel time behavior of the VJs on the analyzed bus line. This knowledge would help understand how the travel time varies depending on the type of calendar day and the time of day and the section of the route studied.

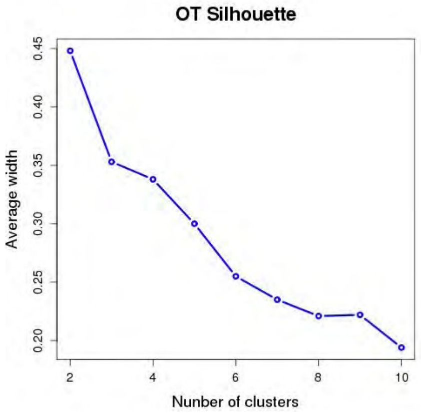


FIGURE 3. Value obtained with the silhouette function for each of the different clusters.

The first step of this phase consisted in modeling TT behavior in 2015. To this end, the k-Medoid clustering technique was applied to the dataset $\{OT\}_{210,2015}$. Nine clustering processes were carried out, generating from 2 to 10 clusters, evaluating in each process the segmentations created with the silhouette function. Fig. 3 contains the average value obtained in each of the nine different cluster groupings, showing that the consistency values decrease as the number of clusters used increases. As an example of the segmentations that were created, Fig. 4 shows the results obtained for the three groupings with the highest value in the metric: the results for two, three and four clusters. The vertical axis represents the arrival time from the start of the VJ, and the horizontal axis the stops analyzed. The three vertical lines represent the three sections into which the route was initially divided: urban, intercity, and urban section. Each graph shows the scheduled time (red line), the medoid of each resulting cluster group (blue line) and the elements classified in each cluster group (gray lines). As may be observed, using two clusters (Fig. 4(a) and 4(b)) the average cohesion value evaluated with the silhouette function is 0.45, and the cohesion values of each of the two clusters are 0.52 (Cluster 1) and 0.36 (Cluster 2). In the case of three clusters, the average value for the silhouette function was 0.35, with the cohesion values for each of the clusters 0.40 (Cluster 1), 0.44 (Cluster 2), and 0.21 (Cluster 3) (see Fig. 4(c), 4(d) and 4(e)). Finally, using four clusters in the group, the average cohesion value for the four clusters was 0.34, the value of each being 0.37 (Cluster 1), 0.36 (Cluster 2), 0.20 (Cluster 3), and 0.33 (Cluster 4) (see Fig. 4(f), 4(g), 4(h) and 4(i)). In this evaluation of the cluster groups, the two groups that produced the highest values were those obtained using two and three clusters. Although the group using two clusters produced the highest average consistency and cohesion value, the group of three clusters gave more precise information about arrival time behavior at the stops. If we compare both groups, we can conclude that the group using three clusters is a refinement of the result obtained with two clusters, and that three TT behavior patterns may be distinguished: Cluster 1 groups the VJs that arrive at the stops in a shorter time, Cluster 2 groups those that take more time than the VJs in Cluster 1, and Cluster 3 groups the VJs with the latest arrival times. In addition, the number of data records in each of the three clusters is significant; 1,777 in Cluster 1; 2,411 in Cluster 2; and 1,904 in Cluster 3. For the above reasons, the grouping of three clusters was taken as the reference for classifying TT behavior.

The second step of the modeling phase consisted of obtaining the behavior patterns for the deviations from the arrival times at the selected stops on the route. To this end, the reference grouping of three clusters was used. Therefore, three patterns were generated, which were defined as the difference function between the observed arrival time and the scheduled arrival time. Fig. 5 shows the data generated with this difference function grouped in each cluster: Cluster 1 Fig. 5(a), Cluster 2 Fig. 5(b) and Cluster 3 Fig. 5(c). The vertical axis represents the deviation of the arrival times from the scheduled times and the horizontal axis, the selected stops. The three vertical red lines represent the same as in Fig. 4. Each graph shows the medoid of the cluster group (blue graphs) and the deviations from the scheduled arrival time for each VJ from each cluster group (gray graphs).

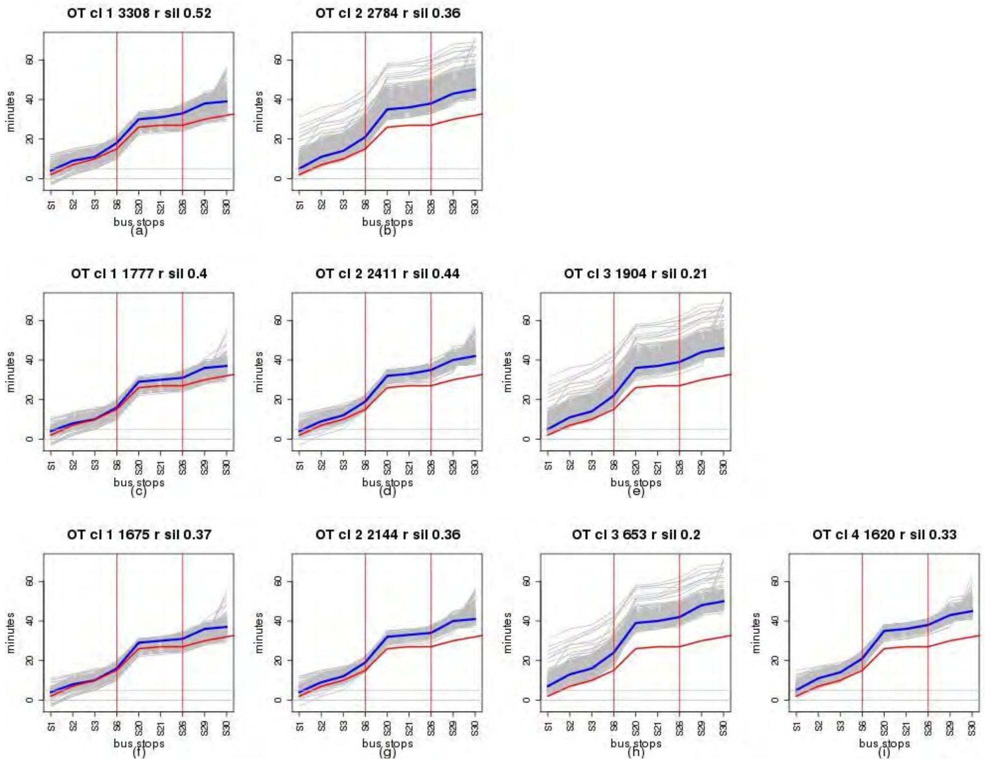


FIGURE 4. Result of clustering the dataset $\{OT\}_{210,2015}$, with two, three and four clusters using the k-Medoid technique.

The third step of the modeling phase identified the sections that generated delay and the sections in which there was a reduction in delays. For this, a group of three clusters was created with the dataset $\{RTD\}_{210,2015}$. The results are shown in Fig. 6. The vertical axis represents the relative deviation of the VJ at each stop, and the horizontal axis the stops analyzed. As in the previous two figures, the three vertical red lines represent the three sections into which the line route was divided. Each graph shows the medoid of each of the three resulting cluster groups (blue graphs) and the elements classified in each cluster group (gray graphs).

V. DISCUSSION

From the results obtained in the analysis of the arrival times at stops, it may be concluded that these times do not follow a single pattern, as was assumed in the bus timetable. From these results three behavior patterns were obtained. The first pattern relates to the VJs that reach the stops on the route in the least amount of time (Cluster 1, Fig. 4(c)), the second pattern, the VJs that take longer than the first cluster (Cluster 2, Fig. 4(d)), and the third, the VJs that take the most time to reach the stops (Cluster 3, Fig. 4(e)). The pattern of Cluster 1—the cluster with the greatest schedule adherence—is represented by its medoid, which indicates a deviation from the schedule that rarely exceeds 5 minutes, the time threshold considered tolerable according to studies carried out by various public transport agencies (see Fig. 5(a)). Nevertheless, it is noteworthy that a considerable number of VJs arrive before the scheduled time (in Fig. 4(c) the VJs below the red line representing the schedule and in Fig. 5(a) with TD the VJs with negative values). Non-adherence with schedules when arriving ahead of time is an event that should not occur on routes that are planned by timetables. Conversely, another behavior evinced by the results is that the greatest VJ delays accumulate on the final part of the route, specifically from stop 20 onwards—the first stop of the county road section as shown in Fig. 2(b). On this final part of the route, delays generally exceed 5 minutes, and in the clusters that show behavior patterns of greater deviation from the schedule, this may even exceed ten minutes. This fact is also relevant to VJ scheduling, since it implies that part of the time planned between the end of the delayed VJ and the next to be carried

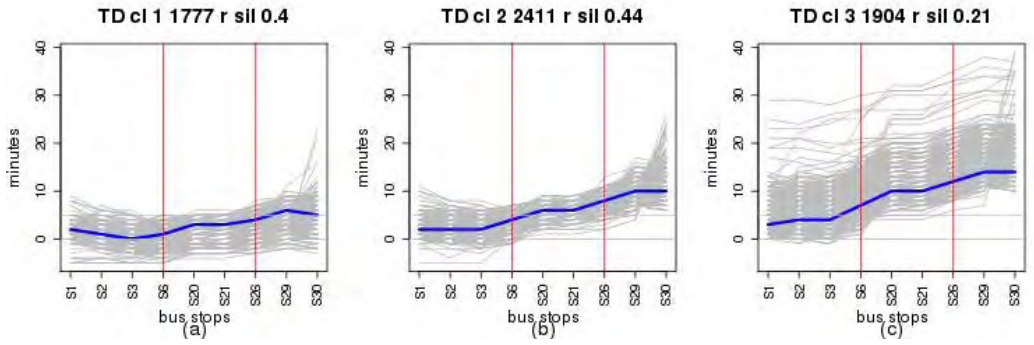


FIGURE 5. Result of the clustering process with 3 clusters applied to the dataset $\{TD\}_{210,2015}$.

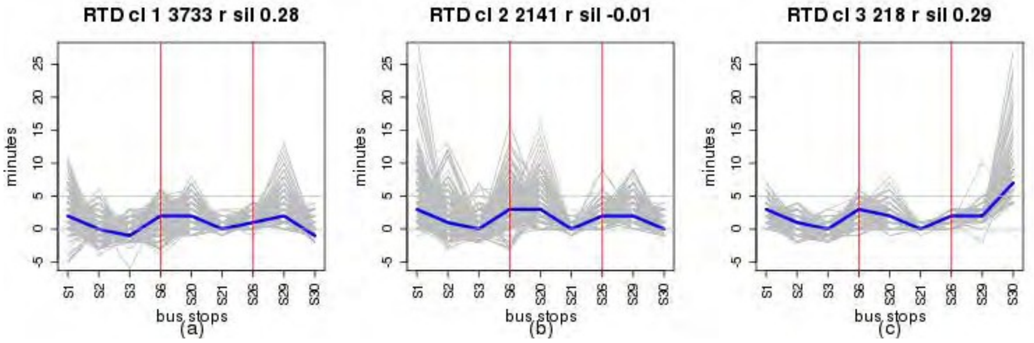


FIGURE 6. Result of the clustering process with 3 clusters applied to the dataset $\{RTD\}_{210,2015}$.

out by the same vehicle—a time interval planned so that the driver can rest and the passengers board the vehicle for its next VJ—is consumed by the delay and may result in the late departure of the next VJ to be made by that vehicle. Finally, another generalized behavior revealed by the results is that the deviations in arrival times at the stops are maintained or increase at the following stops on the route. This behavior can be clearly seen in the forms of the medoids in the three graphs of Fig. 5.

As has already been mentioned, it is clear from the results that the scheduling of arrival times assuming constant values at each stop on the route is not realistic. This statement is supported by the fact that the resulting clusters have a considerable number of samples and their medoids acquire different forms. The question that arises now is how to analyze the relationship between them and the type of day and time of day. To conduct this analysis, contingency tables have been used to represent the frequency with which these patterns occur on different types of day and times of day. Fig. 7 shows these tables for the grouping of three clusters. To analyze the relationship with the type of calendar day, two contingency tables were obtained; one with the months of the year (Fig. 7(b)) and another with the days of the week (Fig. 7(c)). To analyze the time of day, four contingency tables were obtained; one with the time of day at which VJs began on ''Monday to Friday excluding public holidays'' (Fig. 7(d) with VJs between 06:00 and 15:00 and 7(e) with VJs between 16:00 and 22:00); another with the time of day at which VJs began on Saturdays (Fig. 7(f)); and another with the time of day when VJs began on Sundays and public holidays (Fig. 7(g)). In the tables shown in Fig. 7(b) and 7(c) it may be seen that, in the month of August and on Sundays or public holidays, the most frequent pattern is Cluster 1: the VJs that takes the least amount of time to arrive at the stops. In the tables that associate the clusters with the time of day (Fig. 7(d), 7(e), 7(f) and 7(g)) it is clear that arrival time behavior varies depending on the type of day; the behavior is different from Monday to Friday, on Saturdays and on Sundays and public holidays. Moreover, for each of these types of day, the behavior varies according to the time of day. From the results it may be concluded that, in order to adapt a timetable to reality, different forecasts should be used that take into account the following:

- The time of year; August differentiated from the rest of the months of the year.
- The type of day, differentiating Monday to Friday excluding public holidays, Saturdays, and Sundays and public holidays.
- Time of day, differentiating time periods and type of day.

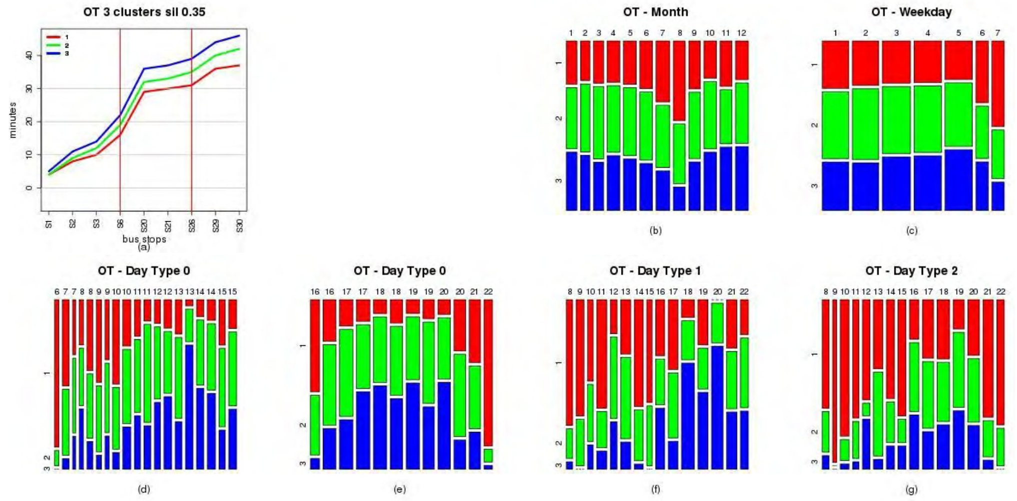


FIGURE 7. Graphs created with the grouping of three OT clusters. (a) Medoids; (b) contingency table of clusters with months of the year; (c) contingency table of clusters with days of the week; (d) contingency table of clusters with VJs on work days (type 0) until 15:00; (e) contingency table of clusters with VJs on work days (type 0) from 16:00 onwards; (f) contingency table of clusters with VJs on Saturdays (type 1); (g) contingency table of clusters with VJs on Sundays and public holidays.

In the analysis of the results obtained in the classifications of the dataset $\{RTD\}_{201,2015}$ formed by the deviations from schedule from one stop to the next on the route, the points of inflection in the medoids are of special interest. These points indicate a change in the behavior of the deviations from the planned schedule, as discussed in Section 3.3, in which the usefulness of this dataset was described. The possible changes are: a section in which the delay decreases; a section in which the delay is maintained; and a section in which the delay increases. In order to improve punctuality, the inflection points marking the beginning of a section in which delays are generated are particularly relevant, since once these sections have been identified, they can be studied to determine the causes of this behavior. At stops 3, 6, 21, 26 and 29 all the medoids have inflection points (see Fig. 6(a), 6(b) and 6(c)). Of these stops, those that begin a section in which a delay is generated are stops 3 and 21 in all the medoids, and stop 29 only in the medoid associated with Cluster 3. The section that begins at stop 3 ends at stop 6, the section that begins at stop 21 ends at stop 26, and the section that begins at stop 29 ends at stop 30. To study the possible causes of this behavior in these sections, it would be necessary to analyze the influences of the DW and RT times on the TT of these sections. If we consider the users of the stops located in these sections, these stops are not the most frequented on the route; this leads to the conclusion that the DW time is not the main cause of the slowness of the vehicle in these sections. To analyze the effect of the RT time on the TT of these routes, the information provided by the transport company's geographic information system was used and it may be observed that a factor that both routes have in common is that they run along single-lane roads in both directions and without any road signs that prioritize public transport vehicles. It could therefore be concluded that the reason for deviations from the schedule in these sections is due to the low speed of the vehicles owing to the conditions of the roads along which they travel. A source of valuable information to analyze the causes of the low speed of these vehicles is GPS readings indicating when vehicles are stationary in these sections, since these readings may follow a pattern that enables these causes to be identified, but this is a subject that falls outside the scope of this paper.

Finally, it should be noted that the proposed methodology enables information on TT behavior to be obtained without specialist knowledge, which would otherwise be necessary if traditional methodologies, based mainly on statistical methods, were used.

VI. CONCLUSIONS

This paper has presented a methodology for analyzing TT in a context of a road-based mass transit system planned by timetables. The methodology, based on data mining, uses the location data of vehicles from the public transport fleet as initial data. It enables the TT of the different scheduled routes to be systematically analyzed, guaranteeing the validity of the results by subjecting the data to validation processes. In addition, in order for the methodology to be suitable for implementation on the greatest possible number of mass transit systems, it has been formalized using standard data models and metrics. From the methodological point of view, the proposal is based on the k-Medoids clustering technique, used to obtain the TT behavior patterns of the VJs, and the silhouette function, used to evaluate the consistency of the clusters.

In the modeling phase, three sets of input data were used. The first dataset, made up of the recorded arrival times at stops, was used to obtain the TT behavior patterns of the analyzed routes. The second dataset, containing the deviations from the scheduled stop times, was analyzed to understand the behavior of these deviations and to detect where the greatest cost is incurred in terms of quality of service. The third dataset, containing the relative deviations in arrival times at each of the stops, was used to obtain information about the TT behavior in the different sections of a route. This information enables the identification of the sections on the route in which scheduled TT deviations occur (late or early arrival). Once these sections have been identified, they may be analyzed individually to detect the places and causes of these deviations.

This paper presents a use case in which the TT of a transport line of a public transport operator was analyzed, using real data provided by the operator. The results have provided information about the TT behavior of this line according to different types of day and times of day. This information enables possible improvements in the scheduling of stops, making it more reliable and thus improving quality of service. It has also made it possible to identify the sections of the route in which the greatest schedule deviations occur.

ACKNOWLEDGMENT

This study was carried out with the collaboration of the public transport company Global Salcai-Utinsa and the Autoridad Única del Transporte de Gran Canaria (Transportation Agency of Gran Canaria)''.

REFERENCES

[1] M. van der Hoeven. *World Energy Outlook*. Int. Energy Agency, Paris, France, Tech. Rep. WEO-2012, 2012, accessed: Feb. 15, 2018. [Online]. Available: https://www.iea.org/publications/freepublications/publication/world-energy-outlook-2012.html

[2] *WHO Releases Country Estimates on Air Pollution Exposure and Health Impact*. World Health Org., Geneva, Switzerland, 2016, accessed: Feb. 15, 2018. [Online]. Available: http://www.who.int/mediacentre/news/releases/2016/air-pollution-estimates/en/

[3] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, ''Data-driven intelligent transportation systems: A survey,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[4] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, ''Traffic flow prediction with big data: A deep learning approach,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[5] V. Guihaire and J. Hao, ''Transit network design and scheduling: A global review,'' *Transp. Res. A, Policy Pract.*, vol. 42, no. 10, pp. 1251–1273, Dec. 2008.

[6] L. Moreira-Matias, J. Mendes-Moreira, J. F. D. Sousa, and J. Gama, ''Improving mass transit operations by using AVL-based systems: A survey,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1636–1653, Aug. 2015.

[7] G. Dimitrakopoulos and P. Demestichas, ''Intelligent transportation systems,'' *IEEE Veh. Technol. Mag.*, vol. 5, no. 1, pp. 74–84, Mar. 2010.

[8] B. Agard, C. Morenc, and M. Trépanier, ''Mining public transport user behaviour from smart card data,'' in *Proc. 12th IFAC Symp. Inf. Control Problems Manuf.*, May 2006, pp. 399–404.

[9] N. Lathia, J. Froehlich, and L. Capra, ''Mining public transport usage for personalised intelligent transport systems,'' in *Proc. 10th IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 887–892.

[10] N. Lathia and L. Capra, ''Mining mobility data to minimise travellers' spending on public transport,'' in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2011, pp. 1181–1189.

[11] N. Lathia, C. Smith, J. Froehlich, and L. Capra, ''Individuals among commuters: Building personalised transport information services from fare collection systems,'' *Pervasive Mobile Comput.*, vol. 9, no. 5, pp. 643–664, Oct. 2013.

[12] B. Du, Y. Y. Yang, and W. Lv, ''Understand group travel behaviors in an urban area using mobility pattern mining,'' in *Proc. IEEE 10th Int. Conf. Ubiquitous Intell. Comput., IEEE 10th Int. Conf. Auto. Trusted Comput.*, Dec. 2013, pp. 127–133.

[13] M. G. Karlaftis and E. I. Vlahogianni, ''Statistical methods versus neural networks in transportation research: Differences, similarities and some insights,'' *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 3, pp. 387–399, Jun. 2011.

[14] R. Xue, D. Sun, and S. Chen, ''Short-term bus passenger demand prediction based on time series model and interactive multiple model approach,'' *Discrete Dyn. Nature Soc.*, vol. 2015, pp. 1–11, Mar. 2015.

[15] D. Celebi, B. Bolat, and D. Bayraktar, ''Light rail passenger demand forecasting by artificial neural networks,'' in *Proc. Int. Conf. Comput., Ind. Eng.*, Jul. 2009, pp. 239–243.

[16] Y. Wei and M. Chen, ''Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks,'' *Transp. Res. C, Emerg. Technol.*, vol. 21, no. 1, pp. 148–162, Apr. 2012.

[17] T.-H. Tsai, C.-K. Lee, and C.-H. Wei, ''Neural network based temporal feature models for short-term railway passenger demand forecasting,'' *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3728–3736, Mar. 2009.

[18] W. Deng, W. Li, and X. Yang, ''A novel hybrid optimization algorithm of computational intelligence techniques for highway passenger volume prediction,'' *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4198–4205, Apr. 2011.

[19] J. Zhao, Q. Qu, F. Zhang, C. Xu, and S. Liu, ''Spatio-temporal analysis of passenger travel patterns in massive smart card data,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3135–3146, Nov. 2017.

[20] N. Uno, F. Karachi, H. Tamura, and Y. Iida, ''Using bus probe data for analysis of travel time variability,'' *J. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 2–15, Feb. 2009.

[21] Y. Bie, X. Gong, and L. Zhiyuan, ''Time of day intervals partition for bus schedule using GPS data,'' *Transp. Res. C, Emerg. Technol.*, vol. 60, pp. 443–456, Nov. 2015.

[22] C. Zhou, P. Dai, and R. Li, ''The passenger demand prediction model on bus networks,'' in *Proc. IEEE 13th Int. Conf. Data Mining Workshops*, Dec. 2013, pp. 1069–1076.

[23] V. T. Tran, P. Eklund, and C. Cook, ''Learning diagnostic diagrams in transport-based data-collection systems,'' in *Foundations of Intelligent Systems* (Lecture Notes in Computer Science), vol. 8502. Cham, Switzerland: Springer, Jun. 2014, pp. 560–566.

[24] F. Pinelli, F. Calabrese, and E. Bouillet, ''A methodology for denoising and generating bus infrastructure data,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 2406–2417, Apr. 2015.

[25] B. Barabino, M. Di Francesco, and S. Mozzoni, ''Rethinking bus punctuality by integrating automatic vehicle location data and passenger patterns,'' *Transp. Res. A, Policy Pract.*, vol. 75, pp. 84–95, May 2015.

[26] S. Mozzoni, R. Murru, and B. Barabino, ''Identifying irregularity sources by automated location vehicle data,'' *Transp. Res. Procedia*, vol. 27, pp. 1179–1186, Sep. 2017.

[27] J. Mendes-Moreira, L. Moreira-Matias, J. Gama, and J. F. de Sousa, ''Validating the coverage of bus schedules: A machine learning approach,'' *Inf. Sci.*, vol. 293, pp. 299–313, Feb. 2015.

[28] J. Khiari, L. Moreira-Matias, V. Cerqueira, and O. Cats, ''Automated setting of bus schedule coverage using unsupervised machine learning,'' in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*, Apr. 2016, pp. 552–564.

[29] R. Jeong and L. Rilett, ''Prediction model of bus arrival time for real-time applications,'' *Transp. Res. Rec.*, vol. 1927, no. 1, pp. 195–204, Jan. 2005.

[30] H. Chang, D. Park, S. Lee, H. Lee, and S. Baek, ''Dynamic multi-interval bus travel time prediction using bus transit data,'' *Transportmetrica*, vol. 6, no. 1, pp. 19–38, Oct. 2009.

[31] W. Lee, W. Si, L. Chen, and M. Chen, ''HTTP: A new framework for bus travel time prediction based on historical trajectories,'' in *Proc. ACM 20th Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2012, pp. 279–288

[32] L. Vanajakshi, S. C. Subramanian, and R. Sivanandan, "Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses," *IET Intell. Transp. Syst.*, vol. 3, no. 1, pp. 1–9, Mar. 2009.

[33] W. Suwardo, M. Napiah, and I. Kamaruddin, "ARIMA models for bus travel time prediction," *J. Inst. Eng.*, vol. 71, no. 2, pp. 49–58, Jan. 2010.

[34] G. Chen, X. Yang, J. An, and D. Zhang, "Bus-arrival-time prediction models: Link-based and section-based," *J. Transp. Eng.*, vol. 138, no. 1, pp. 60–66, Jan. 2012.

[35] B. Yu, Z. Yang, K. Chen, and B. Yu, "Hybrid model for prediction of bus arrival times at next station," *J. Adv. Transp.*, vol. 44, no. 3, pp. 193–204, Jul. 2010.

[36] C. Shearer, "The CRISP-DM model: The new blueprint for data mining," *J. Data Warehousing*, vol. 5, no. 4, pp. 13–23, Jan. 2000.

[37] *The CEN Public Transport—Reference Data Model*, CEN Standard CEN/TR 12896-9:2016, 2016.

[38] N. Paulley *et al.*, "The demand for public transport: The effects of fares, quality of service, income and car ownership," *Transp. Policy*, vol. 13, no. 4, pp. 295–306, Jul. 2006.

[39] J. Strathman, T. Kimpel, and K. Dueker, "Automated bus dispatching, operations control and service reliability," *Transp. Res. Rec.*, vol. 1666, pp. 28–36, Jun. 1999.

[40] M. Dessouky, R. Hall, L. Zhang, and A. Singh, "Real-time control of buses for schedule coordination at a terminal," *Transp. Res. A, Policy Pract.*, vol. 37, no. 2, pp. 145–164, Feb. 2003.

[41] P. G. Furth and T. H. J. Muller, "Service reliability and optimal running time schedules," *Transp. Res. Rec.*, vol. 2034, pp. 55–61, Dec. 2007.

[42] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[43] L. Kaufman and P. J. Rousseeuw, "Partitioning around medoids (program PAM)," in *Finding Groups in Data Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 1990, ch. 2, pp. 68–125.

[44] E. Aldana-Bobadilla and A. Kuri-Morales, "Clustering method based on the maximum entropy principle," *Entropy*, vol. 17, no. 1, pp. 151–180, Jan. 2015.

[45] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.

[46] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. (2017). *Cluster Analysis Basics and Extensions, R Package Version 2.0.6*. [Online]. Available: shttps://CRAN.R-project.org/package=cluster

TERESA CRISTÓBAL received the B.S. degree in computer science and the M.S. degree in master's degree in intelligent systems and numeric applications in engineering from the University of Las Palmas de Gran Canaria, Canary Islands, Spain, in 1990 and 2014, respectively, where she is currently pursuing the Ph.D. degree in enterprise, Internet and communications technologies.

Since 2012, she has been a Research Assistant with the Institute for Cybernetic, University of Las Palmas de Gran Canaria. Her research interest includes the development of intelligent transport systems for public transport and using data mining-based models for public information services. She is the author of eight articles.

GABINO PADRÓN was born in Caracas, Venezuela, in 1966. He received the B.S. and M.S. degrees in computer engineering from the University of Las Palmas de Gran Canaria, Canary Islands, Spain, in 1990, and the Ph.D. degree in computer engineering from the University of Las Palmas de Gran Canaria in 2015.

Since 1990, he has been a Professor with the Informatic and Systems Department, University of Las Palmas de Gran Canaria. He has authored three books and 25 articles. His research interests include passenger transport system focused on transport network planning, AVL systems, and global positioning system.

ALEXIS QUESADA-ARENCIBIA received the B.S. and M.S. degrees in computer science and the Ph.D. degree in computer science from the University of Las Palmas de Gran Canaria (ULPGC) in 1997 and 2001, respectively. He is currently a Doctor-Employed Teacher with the Computer Science and Systems Department, ULPGC, where he is also the Director of the University Institute for Cybernetics.

His main lines of research include the fields of cybernetics, robotics, artificial vision, and intelligent transport systems. He has authored 70 articles in international and national journals, co-authored five books, and edited 14, eight of them in international journals. He is an assessor of different international journal and conferences. He has taken part in over 30 international conferences and has participated in the organization of over 10. He has taken part in seven research projects, being the lead researcher in three of them. He has also participated in six investigation contracts as the lead researcher. Since 2004, he has been teaching Ph.D. course in the Cybernetics and Telecommunication Program, where he has been the Director since 2011. He has directed the development commission of the new doctorate program—Company, Internet and Communications Technologies—in which he currently teaches different activities. He has directed over 50 final degree projects (engineering, bachelor's, and master's degrees), has been part of several Ph.D. examining committees and directed a doctoral thesis; at present five Ph.D. students are under his tutelage. He has directed and has been a speaker in over 60 training courses.

FRANCISCO ALAYÓN was born in Las Palmas de Gran Canaria, Spain, in 1964. He received the B.S. and M.S. degrees from the University of Las Palmas de Gran Canaria, Canary Islands, Spain, in 1989, and the Ph.D. degree from the University of Las Palmas de Gran Canaria in 2007, all in computer engineering.

Since 1989, he has been a Professor with the Informatic and Systems Department, University of Las Palmas de Gran Canaria. He has authored over 50 articles and 20 inventions. He holds one patent. His research interests include passenger transport system focuses in transport network planning, communications systems, integration of the transport vehicle devices in the company's data network.

CARMELO R. GARCÍA was born in Las Palmas de Gran Canaria, Spain, in 1963. He received B.S. and M.S. degrees in Computer Science from the University of Las Palmas de Gran Canaria, Gran Canaria, Spain, in 1989, and the Ph.D. degree in computer science from the University of Las Palmas de Gran Canaria in 1995.

Since 1987, he has been a Professor with the Informatics and Systems Department, University of Las Palmas de Gran Canaria. He has authored one book, over 70 articles, and 20 inventions. He holds one patent. His research interests include ubiquitous computing, intelligent transport systems, and new technologies for education.

3. Conclusiones finales

Los artículos presentados se enmarcan dentro del ámbito de los Sistemas de Transporte Inteligentes para el transporte público de viajeros interurbano por carretera, donde los servicios están planificados por horario, no por frecuencia, por lo que la demanda está condicionada por el servicio [50]. Aunque en algún caso es posible la generalización, la mayor contribución de estos documentos en el ámbito concreto del transporte interurbano, es la propuesta de soluciones sistemáticas para el análisis de la demanda y el control de los tiempos en las operaciones de transporte.

Por otro lado, también cabe destacar que se basan exclusivamente en datos recogidos en la propia actividad, relacionados con el tiempo (hora de salida de la expedición, de llegada y salida a las paradas), el espacio (ubicación de las paradas) y datos generados por sistemas de abordo (de posicionamiento y de pago), por lo que no necesitan de una infraestructura ni de recursos específicos para poder ser llevados a cabo. Así mismo, se han desarrollado con software libre, por lo que facilita la innovación y la transferencia de conocimiento a la empresa.

Es indiscutible el nuevo enfoque que se está dando en los últimos años a la movilidad: las ciudades se están empezando a diseñar para personas, no para automóviles, las autoridades ofrecen subsidios para impulsar modos de transporte sostenibles y las empresas de transporte público personalizan sus servicios. Surgen nuevos modelos como la arquitectura CHIP (Conectada Heterogénea Inteligente Personalizada) para incorporar sistemas de movilidad multimodo basados en conectividad ubicua [51], y para alentar la creación de soluciones innovadoras y creativas, las autoridades y las empresas comienzan a publicar datos sobre la movilidad de los ciudadanos [52].

En este marco aparecen nuevos conceptos y áreas de trabajo, como el de *infomobilidad*, que abarca todo lo relacionado con los sistemas de gestión de la información necesarios para los gerentes del transporte y los viajeros [53], el de *energía informática*, para reconocer el papel que juegan los sistemas de información en el incremento de la eficiencia energética, los *sistemas de información verdes* [54].

Por último, indicar que los trabajos que forman parte de esta tesis por compendio se encuentran dentro de estas líneas de futuro, con la particularidad de que están orientados al transporte interurbano por carretera, cuando muchas de las publicaciones relacionadas con estas nuevas áreas se encuentran enfocadas en el transporte urbano [55].

Referencias

[1] A. Sen, *Development as Freedom*. Oxford University Press, 2001.

[2] WHO, «Road traffic injuries». [En línea]. Disponible en: https://www.who.int/violence_injury_prevention/road_traffic/en/. [Accedido: 28-feb-2019].

[3] WHO, «Air pollution». [En línea]. Disponible en: https://www.who.int/airpollution/en/. [Accedido: 28-feb-2019].

[4] M. Arena, G. Azzone, y S. Malpezzi, «Review on the Infomobility Quality-a new framework».

[5] J. Gubbi, R. Buyya, S. Marusic, y M. Palaniswami, «Internet of Things (IoT): A vision, architectural elements, and future directions», *Futur. Gener. Comput. Syst.*, vol. 29, n.º 7, pp. 1645-1660, 2013.

[6] «Transmodel». [En línea]. Disponible en: http://www.transmodel-cen.eu/overview/use-of-the-transmodel/.

[7] C. Frawley, Wiliam J; Piatetsky-Shapiro, Gregory; Matheus, «Knowledge Discovery in Databases: An Overview», *AI Mag.*, vol. 13, n.º 3, pp. 213-228, 1992.

[8] U. Fayyad, G. Piatetsky-Shapiro, y P. Smyth, «From Data Mining to Knowledge Discovery in Databases», *AI Mag.*, vol. 17, n.º 3, 1996.

[9] M. G. Karlaftis y E. I. Vlahogianni, «Statistical methods versus neural networks in transportation research: Differences, similarities and some insights», *Transp. Res. Part C Emerg. Technol.*, vol. 19, n.º 3, pp. 387-399, nov. 2011.

[10] V. Guihaire y J.-K. Hao, «Transit network design and scheduling: A global review», *Transp. Res. Part A Policy Pract.*, vol. 42, n.º 10, pp. 1251-1273, dic. 2008.

[11] L. Moreira-Matias, J. Mendes-Moreira, J. F. de Sousa, y J. Gama, «Improving Mass Transit Operations by Using AVL-Based Systems: A Survey», *Ieee Trans. Intell. Transp. Syst.*, vol. 16, n.º 4, pp. 1636-1653, ago. 2015.

[12] B. Agard, C. Morency, y M. Trépanier, «Mining public transport user behaviour from smart card data», *IFAC Proc. Vol.*, vol. 12, n.º PART 1, 2006.

[13] N. Lathia, J. Froehlich, y L. Capra, «Mining public transport usage for personalised intelligent transport systems», *Proc. - IEEE Int. Conf. Data Mining, ICDM*, n.º October 2009, pp. 887-892, 2010.

[14] N. Lathia y L. Capra, «Mining Mobility Data to Minimise Travellers' Spending on Public Transport», *Analysis*, pp. 1181-1189, 2011.

[15] N. Lathia, C. Smith, J. Froehlich, y L. Capra, «Individuals among commuters: Building personalised transport information services from fare collection systems», *Pervasive Mob. Comput.*, vol. 9, n.° 5, pp. 643-664, 2013.

[16] B. Du, Y. Yang, y W. Lv, «Understand group travel behaviors in an urban area using mobility pattern mining», *Proc. - IEEE 10th Int. Conf. Ubiquitous Intell. Comput. UIC 2013 IEEE 10th Int. Conf. Auton. Trust. Comput. ATC 2013*, pp. 127-133, 2013.

[17] R. Xue, D. Sun, y S. Chen, «Short-Term Bus Passenger Demand Prediction Based on Time Series Model and Interactive Multiple Model Approach», *Discret. Dyn. Nat. Soc.*, p. 682390, nov. 2015.

[18] D. Celebi, B. Bolat, y D. Bayraktar, *Light Rail Passenger Demand Forecasting by Artificial Neural Networks*. 2009.

[19] Y. Wei y M.-C. Chen, «Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks», *Transp. Res. Part C-Emerging Technol.*, vol. 21, n.° 1, pp. 148-162, abr. 2012.

[20] T.-H. Tsai, C.-K. Lee, y C.-H. Wei, «Neural network based temporal feature models for short-term railway passenger demand forecasting», *Expert Syst. Appl.*, vol. 36, n.° 2, pp. 3728-3736, 2009.

[21] W. Deng, W. Li, y X. Yang, «A novel hybrid optimization algorithm of computational intelligence techniques for highway passenger volume prediction», *Expert Syst. Appl.*, vol. 38, n.° 4, pp. 4198-4205, 2011.

[22] J. Zhao, Q. Qu, F. Zhang, C. Xu, y S. Liu, «Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data», *IEEE Trans. Intell. Transp. Syst.*, vol. 18, n.° 11, pp. 3135-3146, nov. 2017.

[23] N. Uno, F. Kurauchi, H. Tamura, y Y. Iida, «Using Bus Probe Data for Analysis of Travel Time Variability», *J. Intell. Transp. Syst.*, vol. 13, n.° 1, pp. 2-15, 2009.

[24] Y. Bie, X. Gong, y Z. Liu, «Time of day intervals partition for bus schedule using GPS data», *Transp. Res. Part C-Emerging Technol.*, vol. 60, pp. 443-456, nov. 2015.

[25] C. Zhou, P. Dai, y R. Li, «The passenger demand prediction model on bus networks», *Proc. - IEEE 13th Int. Conf. Data Min. Work. ICDMW 2013*, pp. 1069-1076, 2013.

[26] V. T. Tran, P. Eklund, y C. Cook, «Learning Diagnostic Diagrams in Transport-Based Data-Collection Systems BT - Foundations of Intelligent Systems», 2014, pp. 560-566.

[27] F. Pinelli, F. Calabrese, y E. Bouillet, «A Methodology for Denoising and Generating Bus Infrastructure Data», *IEEE Trans. Intell. Transp. Syst.*, vol. 16, n.° 2, pp. 1042-1047, 2015.

[28] B. Barabino, M. Di Francesco, y S. Mozzoni, «Rethinking bus punctuality by integrating Automatic Vehicle Location data and passenger patterns», *Transp. Res. Part A Policy Pract.*, vol. 75, pp. 84-95, may 2015.

[29] S. Mozzoni, R. Murru, y B. Barabino, «Identifying Irregularity Sources by Automated Location Vehicle Data», *Transp. Res. Procedia*, vol. 27, pp. 1179-1186, ene. 2017.

[30] J. Mendes-Moreira, L. Moreira-Matias, J. Gama, y J. Freire de Sousa, «Validating the coverage of bus schedules: A Machine Learning approach», *Inf. Sci. (Ny).*, vol. 293, pp. 299-313, feb. 2015.

[31] J. Khiari, L. Moreira-Matias, V. Cerqueira, y O. Cats, *Automated Setting of Bus Schedule Coverage Using Unsupervised Machine Learning*. 2016.

[32] R. Jeong y L. R Rilett, *Prediction Model of Bus Arrival Time for Real-Time Applications*, vol. 1927. 2005.

[33] H. Chang, D. Park, S. Lee, H. Lee, y S. Baek, «Dynamic multi-interval bus travel time prediction using bus transit data», *Transportmetrica*, vol. 6, n.° 1, pp. 19-38, ene. 2010.

[34] W.-C. Lee, W. Si, L.-J. Chen, y M.-C. Chen, «HTTP: a new framework for bus travel time prediction based on historical trajectories», en *SIGSPATIAL/GIS*, 2012, pp. 279-288.

[35] L. Vanajakshi, S. C. Subramanian, y R. Sivanandan, «Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses», *IET Intell. Transp. Syst.*, vol. 3, n.° 1, pp. 1-9, 2009.

[36] Suwardo, M. Napiah, y I. B. Kamaruddin, «ARIMA MODELS FOR BUS TRAVEL TIME PREDICTION», en *Journal of the Institution of Engineers*, 2010, pp. 49-58.

[37] C. Guojun, Y. Xiaoguang, A. Jian, y Z. Dong, «Bus-Arrival-Time Prediction Models: Link-Based and Section-Based», *J. Transp. Eng.*, vol. 138, n.° 1, pp. 60-66, ene. 2012.

[38] T. Hey, S. Tansley, y K. Tolle, *The FourTh Paradigm. Data-Intensive Scientific Discover*. 2009.

[39] C. Shearer *et al.*, «The CRIS-DM model: The New Blueprint for Data Mining», *J. Data Warehousing14*, vol. 5, n.° 4, pp. 13-22, 2000.

[40] P. J. Rousseeuw, «Silhouettes: A graphical aid to the interpretation and validation of cluster analysis», *J. Comput. Appl. Math.*, vol. 20, pp. 53-65, 1987.

[41] G. Zhang, B. E. Patuwo, y M. Y. Hu, «Forecasting with artificial neural networks: The state of the art», vol. 14, pp. 35-62, 1998.

[42] ORACLE, «SQL Developer». [En línea]. Disponible en: https://www.oracle.com/database/technologies/appdev/sql-developer.html. [Accedido: 01-mar-2019].

[43] Hitachivantara, «Data Integration - Kettle». [En línea]. Disponible en: https://community.hitachivantara.com/docs/DOC-1009855-data-integration-kettle. [Accedido: 01-mar-2019].

[44] «RStudio». [En línea]. Disponible en: https://www.rstudio.com/. [Accedido: 01-mar-2019].

[45] «R Project». [En línea]. Disponible en: https://www.r-project.org/. [Accedido: 01-mar-2019].

[46] «GLOBAL S.U.» [En línea]. Disponible en: https://www.guaguasglobal.com/empresa/magnitudes/. [Accedido: 01-mar-2019].

[47] G. Padrón, F. Alayón, T. Cristóbal, A. Quesada-Arencibia, y C. R. García, *Arrival time estimation system based on massive positioning data of public transport vehicles*, vol. 10070 LNCS. 2016.

[48] C. R. García, A. Quesada-Arencibia, T. Cristóbal, G. Padrón, y F. Alayón, «Systematic development of intelligent systems for public road transport», *Sensors (Switzerland)*, vol. 16, n.º 7, 2016.

[49] T. Cristóbal, J. J. Lorenzo, y C. R. García, «Using data mining to improve the public transport in Gran Canaria island», *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9520, pp. 781-788, 2015.

[50] P. G. Furth y T. H. J. Muller, «Service Reliability and Optimal Running Time Schedules», *Transp. Res. Rec.*, vol. 2034, n.º 1, pp. 55-61, ene. 2007.

[51] V. Sumantran, C. Fine, y D. Gonsalvez, *Faster, Smarter, Greener: The Future of the Car and Urban Mobility*. 2017.

[52] Mayor of London, «London Datastore». [En línea]. Disponible en: https://data.london.gov.uk/. [Accedido: 01-mar-2019].

[53] M. Arena, G. Azzone, F. Franchi, y S. Malpezzi, «Infomobility: a holistic framework for a literature review», *Int. J. Crit. Infrastructures*, vol. 11, n.º 2, pp. 115-135, 2015.

[54] R. T. Watson, M.-C. Boudreau, y A. J. Chen, «Information Systems and Environmentally Sustainable Development: Energy Informatics and New Directions for the IS Community», *MIS Q.*, vol. 34, n.º 1, pp. 23-38, 2010.

[55] R. Wittstock y F. Teuteberg, «Transforming urban public mobility: A systematic literature review and directions for future research», *MKWI 2018 - Multikonferenz Wirtschaftsinformatik*. 2018.